LEZIONE 1

Esame: analisi critica di un lavoro scientifico A NOSTRA SCELTA, che deve essere un lavoro degli ultimi 12 mesi. Cerchiamo di capire come è organizzato il disegno sperimentale. Poi test a risposte multiple come pretest, un esame pratico di 2 ore e un esercizio su R di 40 minuti + un orale facoltativo per alzare il voto, al massimo 3 punti in più.

POSSO FARMI UN FORMULARIO

NELLA PARTE DI R DELL'ESAME CHIEDE DI FARE LE STESSE STATISTICHE FATTE CON MASSOLO ANCHE CON R

VALUTAZIONE

- FASE 1: Analisi critica di un lavoro scientifico (*online* prima dell'esame)
- FASE 2: *Pre-screening* per verifica conoscenze di base
  - Quiz a risposte multiple (*online* in presenza), **40 min**
- FASE 3: ESAME FINALE CON PROVA SCRITTA
  - Esercizio pratico: **2 ore** (su carta in presenza)
  - Esercizio di R: **20 minuti** (piattaforma *online* in presenza)
- ORALE **FACOLTATIVO** (eccetto per chi fa in remoto)(+/- 3 PUNTI; 2 su generale, 1 su R)

DIECI (+ 1) REGOLE PER L'ESAME DI BIOSTATISTICA

LEZIONE 2

La statistica aiuta a raccogliere i dati, organizzare i dati, visualizzare sia graficamente che numericamente, analizzare la variabilità. Si occupa di stime, interpreta il significato dei dati che vengono raccolti.
La statistica è una disciplina che permette di descrivere e misurare diversi aspetti della natura basandosi su campioni. Ma, soprattutto, la statistica consente di **quantificare l'incertezza di queste misure, ovvero di determinare la probabile entità del loro scostamento dal valore vero.**
La biostatistica è la branca della statistica che fa tutte queste cose in ambito biologico. Fisher (quello del test F) è uno dei più importanti biostatistici.

La biostatistica studia le caratteristiche di sistemi biologici: da un insieme di singoli dati permette di passare dal particolare al generale.
POPOLAZIONE STATISTICA: insieme di elementi distinti tra loro di cui si vuole studiare una caratteristica. **Se si prendono tutti gli individui della popolazione e si misura la caratteristica, allora quella caratteristica è il PARAMETRO**. Spesso, però, non possiamo studiare tutta la popolazione. È necessario prendere in considerazione un piccolo gruppo di individui di quella popolazione, il **CAMPIONE**, di cui si misura la caratteristica. In questo caso, si ottiene un **valore diverso da quello del parametro, che è la STIMA**!

La STIMA è il processo con cui si inferisce una grandezza incognita di una popolazione usando dati campionari.
Il PARAMETRO è una grandezza che descrive una popolazione, mentre la STIMA è una grandezza correlata calcolata sulla base del campione.

La statistica si occupa anche della verifica delle ipotesi (o **TEST DELLE IPOTESI**), che è il processo che permette di valutare se un'ipotesi "nulla" che riguarda una popolazione sia compatibile con i risultati osservati. L'**IPOTESI NULLA** è una particolare affermazione sul parametro della popolazione, e rappresenta spesso il punto di vista scientifico, che si vuole verificare o confutare sulla base dei dati raccolti.

Se faccio una STIMA ci sarà un ERRORE rispetto al valore parametrico. **L'errore o INCERTEZZA della stima, la DIFFERENZA RISPETTO AL VALORE PARAMETRICO**, è dovuto non solo alla variabilità biologica, ma anche e soprattutto al **CAMPIONAMENTO: individuazione di un sottoinsieme della popolazione, il campione, di cui si misurano le caratteristiche, e misurandole si ottengono i dati**.
Se l'errore è associato soltanto alla variabilità e casualità con cui vengono selezionati i membri del campione si parla di errore di campionamento.
Campionando si ottiene un valore diverso dal parametro: l'obiettivo ultimo di un disegno sperimentale è campionare in modo **RAPPRESENTATIVO**, fare in modo che la **stima converga verso il parametro**.

Dobbiamo quindi usare degli strumenti che permettano al campione di essere rappresentativo della popolazione. Importante per questo è **scegliere un numero adeguato di soggetti da mettere nel campione**: più n è alto, più la stima diventa simile al parametro. Se i membri del campione sono molte variabili, vuol dire che è meglio prendere un n più alto. Questo è l'**ERRORE CAMPIONARIO**, che emerge dal singolo processo di campionamento ammettendo di aver rispettato tutti i principi del campionamento. Questo errore è quindi dovuto al caso.

Unità statistiche: oggetti che misuro. Le unità statistiche che compongono il campione vengono dette unità CAMPIONARIE.

Quando un campione è RAPPRESENTATIVO: si possono prendere tanti campioni rappresentativi, e ciascuno di loro avrebbe una media diversa da quella del parametro e diversa tra loro. A seconda della domanda di ricerca che ci facciamo, un campione può o meno essere rappresentativo.

Non sempre succede a causa di **misurazioni DISTORTE o BIAS**: se il **campione preso in considerazione è un sottoinsieme distorto allora saranno distorti anche i risultati della misurazione rispetto al parametro reale**. Il bias è una discrepanza sistematica tra le stime e il parametro, che può essere associata ad un campionamento non corretto o ad uno strumento non adeguato, che rendono la stima troppo distante dal valore parametrico.

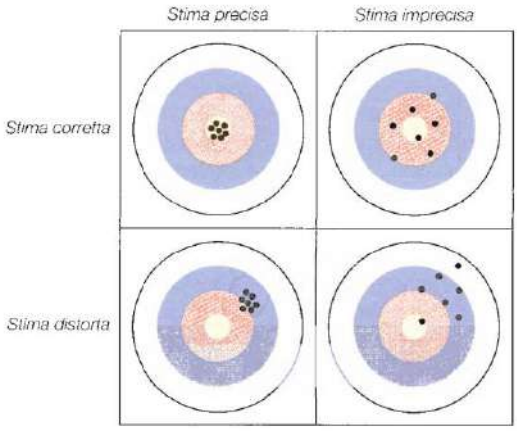
ACCURATEZZA e ADEGUATEZZA della rappresentazione

Cosa vuol dire che un **campione deve essere NON DISTORTO**? La distorsione si ha nel momento in cui si campiona solo una parte della popolazione con caratteristiche particolari, che rendono i dati rilevati dalla misurazione non corretti. Si parla di bias, distorsione. Supponiamo per esempio di dover misurare le caratteristiche dei sepali dei fiori. Se decido per qualche fattore esterno di prendere solo i sepali più colorati, avrò una misurazione distorta. In questo caso, quindi, il bias è associato al ricercatore, ma dipende molto dalla situazione, può essere associato a qualunque cosa. Per **evitare la distorsione posso usare la RANDOMIZZAZIONE**, oppure posso fare una **selezione SISTEMATICA tramite un TRANSETTO** (scelgo ogni tot metri di prato il punto in cui voglio misurare). A seconda di come è la popolazione da campionare posso avere un campionamento random o un campionamento sistematico. Quello random è il più efficace, che riduce al massimo la distorsione.

Un campione DISTORTO è NON RAPPRESENTATIVO.

Campionamento uniforme = sistematico

Supponiamo di avere tre studenti che misurano i sepali. Vengono fuori risultati diversi. Come si fa? Si parte dal confrontare i risultati. C'è il rischio che tutti e tre abbiano la stessa distorsione, e qui il problema è altrove. **Però il confronto tra gli operatori è importantissimo perché, se i risultati differiscono in maniera significativa, vuol dire che il bias dell'operatore è elevato** (magari ha misurato in modo diverso dagli altri)



STIMA

Lo scopo di un buon campionamento è ridurre al minimo l'errore di campionamento e la distorsione delle stime. Sono quindi **necessarie ACCURATEZZA e PRECISIONE**.

Prendo una freccia e cerco di colpire un bersaglio. ogni freccia è un'unità statistica. Non so dove sia il centro del bersaglio, perché non conosco obv le misure. Una **stima CORRETTA e PRECISA è molto vicina al parametro**. Una stima **IMPRECISA è una stima MOLTO VARIABILE**. Una **DISTORSIONE invece sposta la misura lontano dal valore parametrico**. Nel momento in cui ho precisione ma un risultato distorto ho misurazioni tutte simili ma diverse dal valore parametrico. La **distorsione dice quanto il risultato ottenuto dista dal target** (è quella che noi chiamiamo accuratezza)

Se il campione è rappresentativo, e c'è precisione e accuratezza, vado vicino al centro.

Come ottenere un campione casuale (**RANDOMIZZAZIONE**)

1. Creare una lista che includa ogni unità della popolazione di interesse e assegnare un numero ad ogni unità
2. Decidere l'N CAMPIONARIO, numero di unità che verranno campionate
3. Generare N interi casuali
4. Campionare le unità con quei numeri

Variabilità

Se prendo una caratteristica di alcuni oggetti e questa non ha tante variazioni, vuol dire che NON è una variabile. Non serve misurare. Ho di solito un campione con variabilità naturale. Procedure nei protocolli che

aumentano la variabilità più di quanto non lo sia quella biologica: strumentazione, metodologia, operatore. Abbiamo poi la variabilità campionaria, dovuta al fatto che sto prendendo un sottoinsieme del campione.
VARIABILE: caratteristica campionaria che si assume variabile in una popolazione, e che può differire da individuo a individuo. **DATO: misura della variabile** (singola misurazione di ciascuna caratteristica di ciascun soggetto)
Variabili: diverse tipologie: continue, discrete, **qualitative** (o **CATEGORICHE**), **quantitative** (o **NUMERICHE**). A seconda del tipo di variabili che voglio descrivere posso usare un test o un altro. A seconda del tipo di dato posso usare delle statistiche di diverso tipo, descrittive o quantitative.
Unità campionaria = unità statistica = caso = individuo = soggetto = unità su cui si fanno i rilevamenti, sottoinsieme di unità che viene campionata.
Variabile = caratteristica = campo = colonna (ad ogni riga della matrice corrisponde un'unità statistica, ad ogni colonna una variabile)
Misura è molto AMBIGUO, perché vuol dire sia variabile che misurazione. Meglio non usarlo.

Tipi di variabili
Osservo un animale, e cerco di dargli dimensioni, sesso, età. Sono osservazioni che possono essere più o meno precise, soprattutto se osservo da lontano. Guardo anche lo stato, le condizioni fisiche. Non tutte le informazioni obv possono essere prese sempre e contemporaneamente. Nell'evento di osservazione si possono generare una serie di dati.
Variabili **NOMINALI** (es. maschio/femmina, gruppo sanguigno, colore: sono tutte le variabili a cui si dà un nome, sono QUALITATIVE, cioè, determinano una qualità), **ORDINALI** (età), **CONTINUE** (peso). **Nominali sono riordinate come CATEGORIALI o CATEGORICHE, quindi QUALITATIVE.** I **dati categorici NON sono rappresentati da un valore numerico**.
Quelle **ordinali e continue sono QUANTITATIVE**, cioè variabili che posso mettere in ordine crescente.
Tutti gli osservabili che per una certa caratteristica che hanno uno stesso valore sono identici. Se una variabile è ordinale, però, oltre a questa informazione posso anche sapere se una classe viene prima di un'altra (es. se uno è più vecchio di un altro, non solo se non hanno la stessa età).
Come valutare una variabile: delta o rapporto
Temperatura: esempio in cui il riferimento lo abbiamo scelto noi, cioè lo zero. A seconda del riferimento che prendo, il rapporto cambia! Non è vero che tra 10 e 20 gradi abbiamo una temperatura il doppio dell'altra, ma solo differiscono di 10 gradi. in questo caso quindi devo fare il **delta**, non il rapporto, perché **non abbiamo un modo univoco per definire l'unità di temperatura.**
Differenza tra rango e intervallo: un intervallo lo posso ricondurre ad un rango, ma un rango non ad un intervallo. Rango: metto i dati in ordine crescente, la funzione "rango" mi dice il punto in cui si trova uno specifico valore nell'elenco. RANGO=posto occupato in un elenco
Dati numerici/quantitativi continui: possono assumere ogni valore numerico reale possibile
Dati numerici discreti: si presentano in unità indivisibili

NON POSSO FARE UNA MEDIA CON UNA VARIABILE ORDINALE PERCHE' QUANDO ORDINO ASSUMO CHE CI SIA LA STESSA DISTANZA TRA UN VALORE E L'ALTRO DELLA VARIABILE, MA NON è PER FORZA COSI'
SI PUO' TEORICAMENTE FARE, MA SI DEVONO FARE ASSUNZIONI (tipo per l'età, anche se è una variabile ordinale, si può fare la media)

PER DEFINIZIONE UNA VARIABILE ORDINALE è UNA VARIABILE IN CUI I SUOI VALORI POSSONO ESSERE MESSI IN UN CERTO ORDINE
Se ci fosse un ordine intrinseco o una gerarchia nei nomi (ad esempio, ordinati alfabeticamente o in base all'età), allora la variabile potrebbe essere considerata ordinale. Tuttavia, nella maggior parte dei casi, i nomi in un elenco di classe sono considerati nominali.

LEZIONE 3

Creazione tabella di dati: la prima cosa importante è definire le variabili e il loro formato. Per capire il formato di una variabile mi devo chiedere che tipo di variabile sia: è una variabile nominale (aka gli elementi vengono solo classificati in una categoria). Ci sono anche le variabili numeriche e ordinali.
Excel lavora per serie di celle INDIPENDENTI, quindi non è vero che una singola colonna ha dati solo con lo stesso formato.
A, infatti, non è una etichetta modificabile a cui posso attribuire un formato definito, ma rimane così.

Esempio tabella: voto di matematica/voto di biostatistica/numero di volte in cui è stato provato l'esame *cringe*

Devo creare le colonne: i nomi delle colonne stanno nella riga 2, perché la riga 1 è già occupata
La prima colonna sarà qualcosa di identificativo dello studente, quindi il numero di matricola. In questo caso è una variabile nominale, visto che non mi interessa metterla in ordine crescente. Quindi è come se fosse una stringa, non mi interessa il fatto che è un numero.
Seconda colonna: corso di studi. Metto il codice perché, se scrivo tutto per intero rischio l'errore di battitura ogni volta. Una cosa che posso fare è creare la **LOOKUP TABLE (LUT)**, la finestrella con tutte le possibili scelte di lauree magistrali. Molto utile per creare delle matrici di dati. Quindi creo UN ALTRO FOGLIO, con la prima colonna che è l'ID, poi nome per esteso, descrizione e note.
La LUT è una tabella che contiene i valori precompilati per certi input. Quando dunque un input specifico viene dato, viene direttamente fornito l'output dalla LUT.
Terza colonna: corso di laurea triennale. Anche qui potrei avere una LUT con le triennali di Pisa ma non conviene (codice laurea triennale, nome, descrizione, note)
Quarta colonna: voto matematica
Quinta colonna: voto biostatistica
Una volta che mi metto a scrivere le righe sotto, inizio a scrivere un codice e compaiono le opzioni della LUT

Come faccio a campionare gli studenti della classe per avere rappresentatività? Non devo rischiare il bias. Devo dare un numero a tutti ed estrarre a caso. Posso però provare a **STRATIFICARE LA POPOLAZIONE** e accertarmi di stare campionando in modo random/aleatorio/casuale. In questo modo sono sicuro di aver tolto il rischio. L'idea è infatti che anche la laurea triennale di provenienza influenza la formazione.
Un altro possibile bias è scegliere solo le persone che stanno davanti o in fondo, visto che quelli che stanno davanti di solito arrivano a fine corso perché sono più interessati. In questo caso quindi ho inserito una colonna: fila classe. In questo caso ha senso farla NUMERICA, per vedere se tra i fattori che influenzano il voto di statistica c'è anche la fila della classe (che però deve essere una media). Un'altra colonna è la frequenza al corso. Bene io ho come frequenza 3 lezioni visto che smetterò di frequentare ORA. Come **frequenza** posso fare una LUT anche qui, in cui metto 0 (non presente), 1 (<75%), 2 (>75%). Questa è una **variabile ORDINALE, non numerica, perché non ha senso fare una media** per esempio. Essendo una variabile ORDINALE, posso fare come operazione la MODA.

Abbiamo quindi la nostra tabella. Ci aiuta molto a capire la natura delle variabili.
Facciamo quindi una tabella così.
Posso fare poi la media dei voti di matematica. Ma è un grande problema, perché la riga non è una unità statistica.

- Creare un file di dati (Utilizzando Excel o altri strumenti analoghi) per archiviare un *set* di dati di vostro interesse o invenzione, che abbia le seguenti componenti:
  - Tabella dati con almeno 10 casi e 5 colonne
  - LUT
  - Descrizione variabili
- Evitare gli errori piu' comuni
  - Formati non omogenei all'interno della stessa variabile
  - Codici non standard
  - Inserire troppe info all'interno della stessa variabile
  - Valori mancanti indicati con simboli strani e non omogenei o coerenti con il formato della variabile (p.es. non mettere 0 quando non hai il dato)

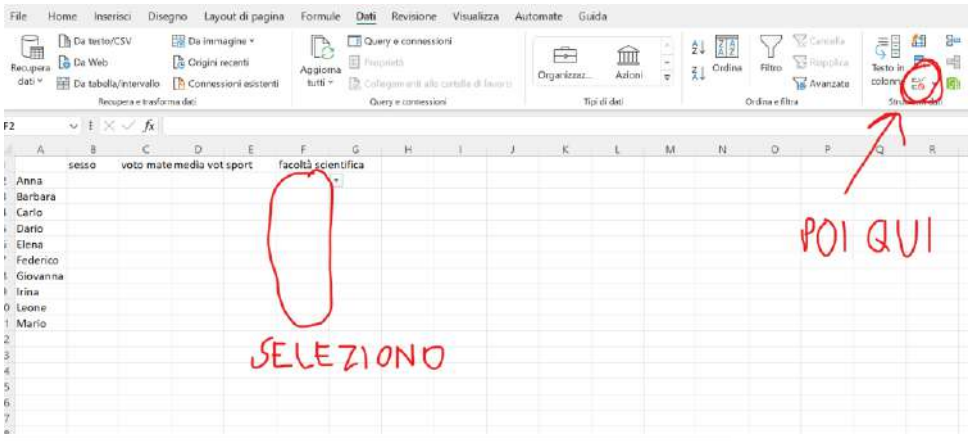
COME CREARE UNA LUT SU EXCEL

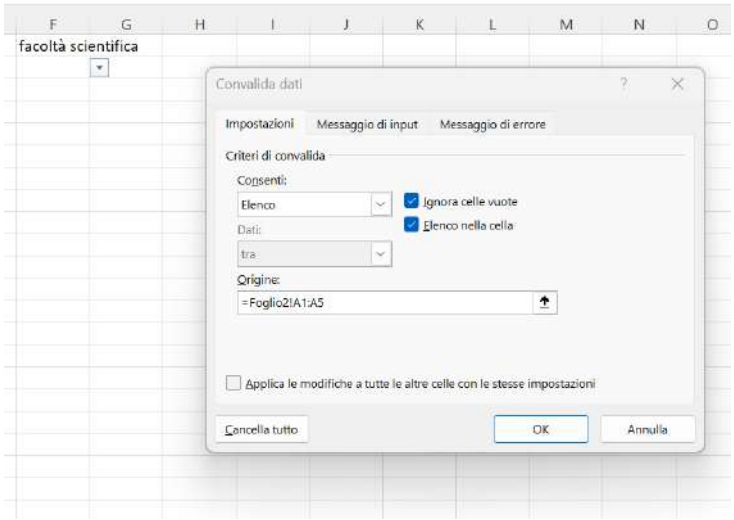
Intanto una LUT è una tabella di ricerca in cui si associano a valori di input dei valori di output. Servono per mappare set di valori in altri set di valori.

1. Creo un nuovo foglio di lavoro in cui faccio la colonna con le opzioni da cui voglio si possa cercare
2. Nel secondo foglio di lavoro, ordina i nomi che hai scritto in ordine alfabetico (o numerico)

	A	B
1	Biologia	
2	Chimica	
3	Fisica	
4	Ingegneria	
5	Matematica	
6		
7		
8		

3. Vai di nuovo sulla tabella principale e seleziona la colonna da cui vuoi compaia la tendina
4. Vai su Dati → Validazione dati → Elenco → Criteri
5. Su Origine scrivi prima il nome del secondo foglio, poi ! e poi le colonne che vuoi siano usate nel menu a tendina





Riprendiamo l'entusiasmante lezione
RAPPRESENTAZIONE DEL DATO VEDI lez0
Non mettere troppi numeri nei dati. Renderli il più chiaro possibile. **Non fare rappresentazione in 3D o grafici a torta. Pensate bene al grafico migliore**. **Non basarsi solo sul colore** (perché tutti non vedono il colore).
Skills comunicative: usare elenchi puntati, grassetto e corsivo
Descrizione variabile continua: migliore rappresentazione grafica quella solita, con punto e barrette perpendicolari (istogramma)

Rappresentazione dei dati nominali – conteggi
Numero di decessi tra gli adolescenti. **Tabella riassuntiva di un dataset:** tabella con causa decesso-numero morti). Il dataset avrà invece molte più informazioni (es. età, luogo di provenienza). Al posto di usare numeri, molto meglio fare una rappresentazione grafica (istogramma solito: GRAFICO A BARRE): con la rappresentazione grafica posso subito vedere il RAPPORTO. Posso usare il grafico a barre per le variabili NOMINALI. Per quelle continue meglio di no, comunque rappresenterebbe una media. In questo grafico sto rappresentando come variabile la causa di morte, che è nominale (aka l'unica cosa che posso fare è contare). Grafico a torta FA SCHIFO. Però magra consolazione dà le percentuali che sono importanti <3

[posso solo contare i dati nominali, e quindi parlo solo di frequenze assolute e relative]

Tabelle di contingenza
Cinciallegra: lo sforzo riproduttivo aumenta la probabilità di sviluppare la malaria? Tolgo/aggiungo le uova e vedo quanti hanno sviluppato la malaria. Prendo solo le femmine, perché i maschi visitano più nidi e non ha senso paragonare maschi e femmine. Due variabili: trattamento e malaria. L'unità statistica è il singolo uccellino. La variabile è malaria: sì/no, trattato: sì/no. Sono tutte e due nominali, e posso contarle e stimare frequenze assolute e relative. Problema: prima avevo una sola variabile. Faccio una **TABELLA DI CONTINGENZA 2x2**: trattati/controlli, malaria/non malaria. Togliendo le uova aumenti lo sforzo di deposizione perché ne depongono di più
Grafico a barre con frequenza assoluta. Si può anche usare il **MOSAIC PLOT**: **l'asse di riferimento va da 0 a 1 (frequenza relativa**). SI USA CON DUE GRUPPI (ES. TRATTATO E CONTROLLO). La larghezza delle colonne è diversa: ci sono più controlli. **Il fatto che anche la larghezza della colonna cambia è importante per l'ATTENDIBILITÀ DELLA STIMA**: un campione è tanto rappresentativo quanto è grande, perché c'è maggiore convergenza verso il parametro reale. Se uno fosse molto largo e l'altro molto stretto non va bene, perché uno dà una stima attendibile e l'altro meno.
[al prof garba tanto il mosaic plot]
[frequenza relativa: estremo è sempre 1]

Grafici a barre
Istogrammi: non sono veri e propri grafici a barre (secondo me NELL'ISTOGRAMMA ABBIAMO una dipendenza tra le variabili (indipendente X e dipendente y, entrambe numeriche) NUMERI SIA IN X CHE IN Y, NEL GRAFICO A BARRE ABBIAMO CARATTERISTICHE/GRUPPI IN X E NUMERI/FREQUENZE IN Y). Sono gli istogrammi di frequenza assoluta. **Capisco che è a frequenza assoluta per la barra verticale, che non arriva né a 1 né a 100**. L'etichetta della barra verticale dice "frequenza come numero di specie", mentre in orizzontale ho l'"abbondanza". Il grafico istogramma rappresenta quante specie hanno lo stesso numero di individui (aka quante specie hanno meno di 50 individui, quante tra 50 e 100, bla bla bla
Grafico di presenza cumulativa: in orizzontale l'abbondanza per specie, in verticale la frequenza relativa cumulativa (aka una sorta di funzione di ripartizione ma non discreta). Sommo ogni volta le specie che considero.

LEZIONE 4

La **FREQUENZA** di una particolare misura in un campione è il numero di volte con cui è stata osservata quella particolare misura. Da qui si definisce la DISTRIBUZIONE DI FREQUENZA, che è la visualizzazione di tutte le frequenze. La frequenza RELATIVA è il rapporto tra la frequenza assoluta e il numero totale di osservazioni

Una **TABELLA DI FREQUENZA** è una visualizzazione del NUMERO DI OCCORRENZE di ogni categoria di un certo data set.si può rendere visivamente più sensata mettendo in ordine di frequenza le occorrenze.

Tabella 2.1-1

Tabella di frequenza che mostra le 10 cause di morte più comuni degli adolescenti statunitensi di età compresa tra 15 e 19 anni nel 1999.

Causa di morte	Frequenza
Incidenti	6688
Omicidio	2093
Suicidio	1615
Tumore maligno	745
Cardiopatia	463
Anomalie congenite	222
Malattia respiratoria cronica	107
Influenza e polmonite	73
Malattie cerebrovascolari	6
Altri tumori	52
Tutte le altre cause	1653
Totale	13 778

Un **DIAGRAMMA A BARRE** usa l'altezza di barre rettangolari per visualizzare la distribuzione di frequenza di una variabile categorica: con il diagramma a barre non si riesce più a visualizzare l'esatta frequenza, ma, in compenso, si ha un quadro chiaro e immediato di come i valori numerici varino in base alla categoria.

REGOLE PER COSTRUIRE UN DIAGRAMMA A BARRE

- La frequenza è data dalla posizione del lato superiore del rettangolo
- È importantissimo che la base inferiore indichi y=0, poi il resto può essere nella scala (coerente) che si vuole o una frequenza relativa (in quel caso arriva fino a 1) o assoluta
- Nel caso di una variabile categorica ordinale, i valori lungo l'asse x dovrebbero essere ordinati in modo naturale (es. da poco a molto), mentre se è nominale è più facile ordinare le varie categorie sulla base della frequenza con cui si osservano, cosa che facilita l'analisi.
- I rettangoli devono essere di uguale larghezza e ciascuno deve poggiare su una linea di base a y=0
- È sempre bene indicare il numero totale di osservazioni (N campionario) nella didascalia della figura.

ISTOGRAMMA

Una delle caratteristiche dell'istogramma è che è un grafico a barre che descrive delle variabili CONTINUE, non delle caratteristiche nominali (aka è una **DISTRIBUZIONE DI FREQUENZA PER UNA VARIABILE NUMERICA**). Le variabili continue sono sull'ascissa, mentre sull'ordinata c'è la frequenza. Se sull'ascissa ho la frequenza **relativa**, se sommo tutti i pezzettini del grafico ottengo 1. Se invece ho la frequenza **assoluta**, ottengo una stima del numero totale di casi. I valori dei dati vengono suddivisi in **intervalli consecutivi sulle x**, o **BIN**, di uguale ampiezza.

L'istogramma deve essere inteso come una **rappresentazione di AREE, NON DI ALTEZZE**! Questo perché può capitare di raggruppare tra loro due bin. In questo caso non si può valutare l'altezza, se no la rappresentazione sarebbe distorta. Oppure sulle y si potrebbe rappresentare la **DENSITA'**, cioè la frequenza divisa per l'ampiezza dell'intervallo. **In questo modo l'area dei rettangoli non è solo proporzionale alla frequenza ma proprio uguale** (la base è l'ampiezza dell'intervallo e l'altezza è la densità)

La **SUPERFICIE TOTALE** **di un istogramma di frequenza assoluta è l'N CAMPIONARIO (secondo me solo se uso la DENSITA' sulle y);** (questo nel caso in cui io stia mettendo nel grafico tutti i casi possibili che considero). Faccio **l'INTEGRALE sotto la curva. In caso sia di frequenza relativa, invece, ho integrale TOTALE 1** (obv integrale se la distribuzione è continua)
Gli istogrammi sono fondamentali per definire la **FORMA DELLA DISTRIBUZIONE DI FREQUENZA**: sono la rappresentazione migliore in statistica per descrivere le distribuzioni. Un intervallo della distribuzione di frequenza più frequente di quelli circostanti è detto **PICCO**. La **MODA** è invece l'intervallo corrispondente al PICCO PIU' ALTO. Per esempio, una distribuzione BIMODALE ha due picchi ma una sola moda, che coincide con il picco più alto.

- Distribuzione **UNIFORME**: bene o male l'altezza di ogni barra è la stessa. Non ci sono valori che prevalgono nettamente rispetto agli altri
- Distribuzione A **CAMPANA**: una delle più importanti in statistica
- Distribuzione a **CAMPANA ASIMMETRICA.** Una distribuzione di frequenza per una variabile numerica può essere se non simmetrica asimmetrica positiva (a destra) o negativa (a sinistra), a seconda della posizione della coda più lunga.
- Distribuzione **BIMODALE**: tipica per due gruppi diversi nella stessa popolazione. Es. in molti casi, se ho una popolazione con maschi e femmine che campiono insieme. Informativa perché suggerisce cosa succede.

Da un istogramma si può arrivare al GRAFICO DI FREQUENZA CUMULATIVA

Valori estremi molto distanti dal resto delle osservazioni sono gli **OUTLIER**. Sono comuni nei dati biologici e possono essere associati ad errori commessi nella registrazione dei dati, ma anche possono essere delle caratteristiche reali nella natura.

L'AMPIEZZA degli intervalli influenza la forma dell'istogramma.

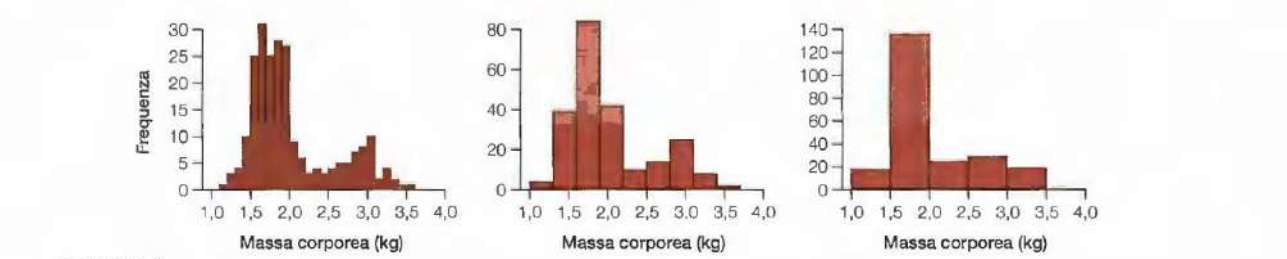


Figura 2.1-4
Massa corporea di 228 femmine di salmone rosso campionate dal Pick Creek in Alaska (Hendry et al., 1999). In ciascun caso sono presentati gli stessi dati, ma le ampiezze degli intervalli sono diverse: 0,1 kg (a sinistra), 0,3 kg (in centro) e 0,5 kg (a destra).

Esempio 2.1C

Quanti picchi?

Il salmone rosso (*Oncorhynchus nerka*) trascorre anni nell'Oceano Pacifico prima di ritornare nel fiume dove è nato per riprodursi e poi morire. Alcuni individui, però, trascorrono più tempo in mare rispetto agli altri. Ogni anno, quindi, i salmoni che si riproducono in un dato fiume possono appartenere a differenti classi di età. Di conseguenza, gli adulti che si riproducono variano per la loro massa corporea. La Figura 2.1-4 mostra tre differenti istogrammi che presentano la massa corporea di 228 femmine di salmone rosso campionate nel Pick Creek, in Alaska, nel 1996 (Hendry et al., 1999).

Caso dei salmoni.: grafici di frequenza relativa. Distribuzione BIMODALE (aka A DUE MODE, con MODA valore più frequente). Come si vede dall'immagine, **più aumento la larghezza delle barre più il dato si disperde, e non compare il secondo picco**. La **larghezza delle barre più piccola implica che ci sia una**

risoluzione maggiore. Se ho **TANTI RECORD** (aka N campionario grande) posso **permettermi di avere un grafico con barre strette**. **BIN PICCOLO vuol dire che ho un grafico che si avvicina a quello vero**.
IL BIN è SULLE X!!! Dividi il tuo range di intervalli in BIN, e a partire dal singolo BIN viene disegnata una barra in verticale che corrisponde alla frequenza del BIN.

Altro esempio: fringuello e spessore del becco. Distribuzione a campana più o meno simmetrica. Posso approssimare alla distribuzione CONTINUA a campana (aka gaussiana o normale). Se campiono lo stesso numero di fringuelli un'altra volta dalla stessa popolazione otterrei un grafico diverso, ma se l'N campionario è RAPPRESENTATIVO, ottengo un grafico simile, che comunque può essere approssimato bene o male alla stessa curva a campana. Grafico di frequenza relativa.

Ma come faccio a capire quante classi mi servono? **BIN SIZE** (CLASSI = BIN)
Obv il numero di classi che mi serve dipende dall'N campionario. I **BIN devono essere tutti identici lungo** l'asse. Se trovo l'ampiezza delle classi vuol dire che so anche le classi, perché so l'ampiezza totale.
Per esempio, nel caso dei salmoni è necessario capire se la distribuzione bimodale sia quella corretta, cioè se ha senso che compaiano due picchi: in questo caso, le due classi di massa corporea corrispondenti ai due picchi erano anche due classi di età, e quindi era sensato che fossero diverse.

REGOLE PER FARE UN ISTOGRAMMA

- Il lato inferiore del rettangolo poggia su y=0.
- I rettangoli vicini di un istogramma sono attaccati l'uno all'altro.
- Per capire il **numero di rettangoli a volte si usa la regola empirica di Sturges, per cui il numero di intervalli è 1+ln(n)/ln(2)**, arrotondando al numero più alto. Nel libro usano intervalli leggermente maggiori.
- È sempre buona regola indicare l'n campionario

x-esimo **PERCENTILE**: valore al di sotto del quale stanno l'x% degli individui. Il **QUANTILE** è invece x/100. Per esempio, il 50esimo percentile è anche il quantile 0,5.

Deviazione standard: misura di dispersione, indicatore delle misure che si ottengono. Dice quanto un valore devia rispetto alla media. Supponiamo per esempio di avere 100 unità statistiche: calcolo la media e vedo quanto scarto ho dalla media. La somma degli scarti dalla media di ogni unità è ZERO. **Varianza**: **somma degli scarti al quadrato.** Ha il problema che aumenta all'aumentare dello scarto, ma anche dell'N campionario. Quindi devo calcolare la **MEDIA DEI QUADRATI COSTRUITI SUGLI SCARTI**. A questo punto sto ancora lavorando sui quadrati, quindi faccio la **RADICE, cioè lo SCARTO QUADRATICO MEDIO, cioè la deviazione standard**

La grandezza della classe dipende da N e dalla deviazione standard: quindi è il rapporto tra variabilità e N campionario che capisco la larghezza della classe (aka h). se la variabilità è bassa, la larghezza può essere piccola, e viceversa. Un campione descrive meglio una popolazione non solo se N è più grande, dipende dalla larghezza della popolazione.

Tabella di contingenza: tabella di frequenza per due o più variabili categoriche che mostra come i valori della prima variabile siano associati ai valori della seconda variabile.
Esempio in questa tabella: i valori della variabile assenza/presenza della malaria sono associate ai valori della seconda variabile controllo/assenza di uova

Tabella 2.3-1

Tabella di contingenza che mostra l'incidenza della malaria aviaria nelle femmine di cinciallegra in relazione al trattamento sperimentale.

	Gruppo di trattamento sperimentale		
	Gruppo di controllo	Gruppo a cui sono state sottratte due uova	Totale delle righe
Malaria	7	15	22
Assenza di malaria	28	15	43
Totale delle colonne	35	30	65

DIAGRAMMA A BARRE RAGGRUPPATE

Usa i rettangoli per rappresentare graficamente tutte le combinazioni di due o più variabili categoriche. La differenza rispetto ai diagrammi a barre è che i gruppi sono rappresentati con diversi colori o diverse sfumature di colore. Lo spazio tra i rettangoli di diversi gruppi deve essere più ampio di quello tra i rettangoli della variabile di risposta.

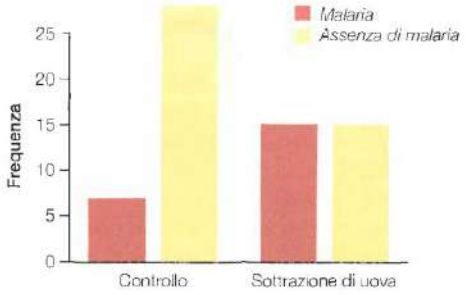


Figura 2.3-1
Diagramma a barre raggruppate sulla relazione fra lo sforzo riproduttivo e la malaria aviaria nelle cinciallegre. I dati sono tratti dalla Tabella 2.3-1, dove $n = 65$ è il numero totale di uccelli.

MOSAIC PLOT

Grafico simile al diagramma a barre raggruppate, con la differenza che i rettangoli che si riferiscono allo stesso gruppo sono impilati l'uno sull'altro invece di stare l'uno accanto all'altro. L'area dei rettangoli rappresenta la **FREQUENZA RELATIVA** entro ciascun gruppo. Anche la **LARGHEZZA** di ogni pila verticale cambia, ed è direttamente proporzionale al NUMERO DI OSSERVAZIONI per quel gruppo. Un grafico a mosaico fornisce solo la frequenza relativa, NON assoluta, si occorrenza di ciascuna combinazione di variabili. Questo potrebbe essere considerato un inconveniente, ma si deve tenere conto che è molto più importante vedere il pattern nei dati che i dati numerici esatti.

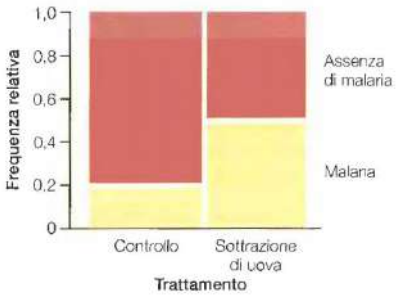


Figura 2.3-2
Grafico a mosaico sulla relazione fra lo sforzo riproduttivo e l'incidenza della malaria aviaria nelle cinciallegre. Il colore giallo indica uccelli affetti da malaria, mentre il colore rosso indica uccelli non affetti da malaria. I dati sono tratti dalla Tabella 2.3-1, dove n = 65 è il numero totale di uccelli.

Quale diagramma è meglio usare? Dipende. Il diagramma a barre raggruppate rispetto alla tabella di contingenza è più facile da visualizzare visivamente: è più facile confrontare l'area o l'altezza rispetto a dei numeri. La complessità del diagramma diventa però troppo alta quando le variabili di risposta hanno molte categorie: in questo caso è meglio usare la tabella, se no ci si perde. Il grafico a mosaico è poi più semplice da leggere del diagramma a barre raggruppate.

CONFRONTO DI ISTOGRAMMI TRA GRUPPI

Livelli emoglobina in maschi umani a diverse altitudini: frequenza relativa. Non aumentano con l'altitudine. Le tre distribuzioni sono simili. Nelle Ande però esiste una popolazione con livelli di emoglobina superiori a quelli degli altri. Posso mettere insieme le frequenze cumulative di tutte le popolazioni campionate.

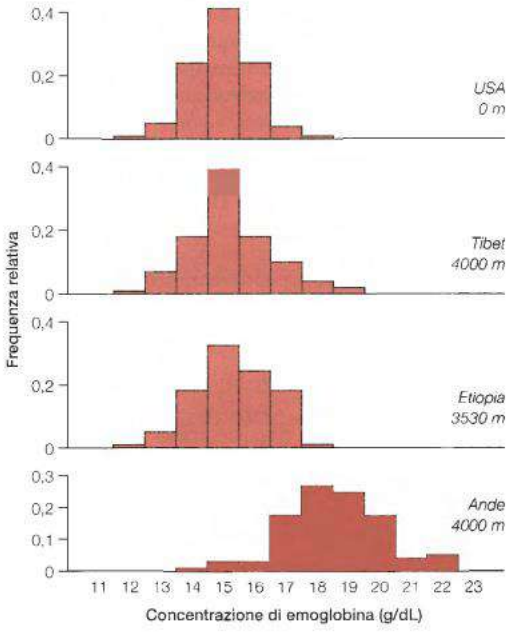


Figura 2.4-1
Istogrammi che mostrano la concentrazione di emoglobina in maschi di popolazioni umane viventi che vivono a elevate altitudini in tre differenti parti del mondo. È inclusa come controllo una quarta popolazione che vive a livello del mare (USA).

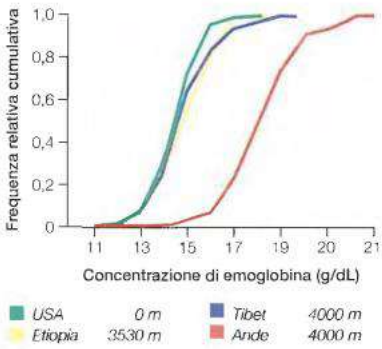


Figura 2.4-2
Distribuzioni di frequenza cumulative della concentrazione di emoglobina in maschi umani che vivono ad altitudini elevate in Etiopia, in Tibet e sulle Ande. È inclusa come controllo una quarta popolazione che vive a livello del mare negli Stati Uniti. (Da Beall et al., 2002; ridisegnato.)

Qui sto analizzando **due variabili: area geografica e livelli di emoglobina**. **Unità statistiche: singoli soggetti su cui ho misurato le due variabili,** emoglobina e il posto da cui viene il soggetto.
Quando devo rappresentare due variabili continue, metto sulle y quella che dipende dall'altra. Per esempio, qui metto l'ALTITUDINE, perché è il variare dell'altitudine che influenza la concentrazione di emoglobina.
Grafico sopra: confronto tra istogrammi
Grafico sotto: confronto tra frequenze cumulative. Le **frequenze cumulative** sono una sorta di funzione di RIPARTIZIONE, che indicano le frequenze di ciascun gruppo, dove le frequenze del gruppo precedente si sommano al successivo, fino ad arrivare a 1.
Il confronto tra frequenze cumulative evidenzia subito la differenza tra le varie aree.

Visualizzare la relazione tra una coppia di variabili numeriche

SCATTER PLOT: diagramma di **dispersione**: OGNI OSSERVAZIONE è UN PUNTO SUL DIAGRAMMA: la posizione delle x del diagramma è la variabile INDIPENDENTE, esplicativa, mentre la posizione sulle y è la variabile DIPENDENTE, di risposta.
Il DIAGRAMMA A LINEE è invece uguale al diagramma di dispersione, ma i punti adiacenti su x sono unii da un segmento. per ogni x quindi si ha una sola y. Questo rende più visibili le tendenze delle popolazioni nel corso del tempo.
Esempio: immagine sotto: associazione temporale: studio equazione di Lotka-Volterra. Sulle ascisse il tempo, sulle ordinate il numero di prede o di predatori. Ci si chiede se c'è ciclicità. L'unità statistica sono due diverse, una la singola preda e una il singolo predatore. Posso anche prendere una singola unità statistica, cioè i tag che i trapper devono lasciare all'associazione quando abbattono un animale e ne prendono la pelle.

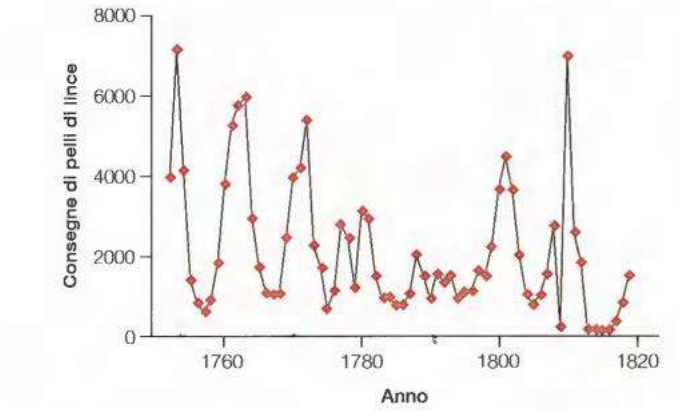


Figura 2.5-2
Diagramma a linee che mostra le consegne di pelli di lince della Hudson's Bay Company dal 1752 al 1819.

Per avere un grafico più informativo è meglio usare il grafico a punti che l'istogramma, perché l'istogramma potrebbe assumere tutti i valori possibili in un bin individuato. Se invece uso sia i puntini che l'istogramma ottengo la rappresentazione migliore possibile, aggiungendo poi le BARRE DI ERRORE: indicano il range dell'errore. **Se uso SOLO istogramma e barre di errore non ve bene, devo aggiungere anche i puntini**.

[IN ASCISSA LA VARIABILE INDIPENDENTE]

Mappe di rappresentazione dati: equivalente spaziale del diagramma a linee: visualizza una misura numerica di risposta in più posizioni su una superficie

REGOLE PRINCIPALI PER UNA VISUALIZZAZIONE EFFICACE DEI DATI

- LESS IS MORE: meno cose si scrivono nei grafici, meglio è
- Mostrare i dati: non mettere soltanto la curva che rappresenta la relazione o l'istogramma, ma anche tutti i puntini che rappresentano la singola unità campionata.
- Rappresentare accuratamente le grandezze: mettere in proporzione le grandezze, per fare in modo che l'area dei rettangoli sia direttamente proporzionale al valore numerico che viene visualizzato: se non si fa così, si amplificano delle differenze magari minimamente significative, facendo falsa informazione.
- Nei diagrammi a barre e negli istogrammi mettere all'origine y=0
- Evitare di includere nel diagramma troppe informazioni, ma anche troppo poche non va bene.
- Fare in modo che la rappresentazione grafica sia di facile interpretazione: scelta dei caratteri, della posizione delle scritte, delle informazioni che si vogliono dare. Non fare abbellimenti non richiesti, scegliere colori appropriati, indicare le unità di misura e i simboli utilizzati.

Le regole per le tabelle sono simili, soprattutto per le tabelle di frequenza, che possono essere sostitutive ad un grafico.

I DIAGRAMMI sono il mezzo migliore per mostrare pattern ed eccezioni, ma non danno molte informazioni quantitative precise. Le tabelle servono quindi quando è importante comunicare aspetti quantitativi dei dati.

Esempi di grafici sbagliati: no cose troppo confuse, mi raccomando ridondanza, puntini e non solo cose continue, no 3D, unità di misura, didascalia, partire da zero (DEVE PARTIRE DA ZERO: se non parte da zero vedo solo il delta e il grafico diventa fuorviante, perchè vedo dei grandi aumenti dove non ci sono).
Il mosaic plot è il suo preferito: meglio B&W

GUARDA BENE L'ULTIMA PARTE DELLE SLIDES lez0

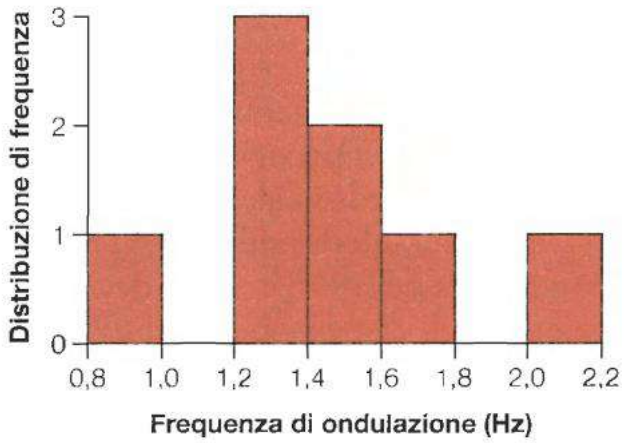
LEZIONE 5

Distribuzioni

Gli istogrammi danno una indicazione dell'andamento di una distribuzione. **Come faccio a sintetizzare la rappresentazione grafica con un numero**? Mi serve una misura centrale, e una dispersione

Una distribuzione viene descritta numericamente da una TENDENZA CENTRALE e una DISPERSIONE

Esempio: frequenze di oscillazione del serpente che, quando si lancia dai rami oscilla, appiattisce il corpo ovunque tranne che nella regione del cuore, e a mano a mano che acquista velocità, assume una forma ad S stretta e orizzontale, e poi fa delle ondulazioni trasversali. In questo modo può scendere di volo pianato dall'albero da cui si è lanciato. Se orienta la testa e la parte anteriore del corpo, riesce a variare la direzione durante il volo ed evitare gli alberi, raggiungendo il punto che vuole.



N è l'integrale (MA PER UN CAZZO, N o è la somma delle altezze in questo caso oppure è l'integrale nel caso in cui i BIN siano 1 oppure nel caso in cui le y rappresentino la DENSITA'), area rossa dell'istogramma. Troviamo che N=8.

La **MEDIA CAMPIONARIA** è la media delle misure del campione, cioè la somma di tutte le misurazioni divisa per n, il numero totale di misurazioni. La **DEVIAZIONE STANDARD** indica quanto distano le misure dalla media, ed è grande se le osservazioni sono lontane dalla media, piccola per il viceversa. La deviazione standard si misura a partire dalla **VARIANZA**, che, come abbiamo detto, è lo SCARTO QUADRATICO MEDIO. Lo SCARTO o DEVIAZIONE è sempre la differenza tra una misura e la sua media. la **DEVIANZA** è la somma degli scarti quadrati. Il **DENOMINATORE della VARIANZA è n-1**!!!! Questo perché si è visto che si ottengono delle stime più corrette sulla varianza della popolazione. L'unità di misura della varianza è il quadrato dell'unità di misura dei valori originali.

La deviazione standard è molto utilizzata, perché indica quanto le misure si discostano dalla media. la deviazione standard è legata alla frequenza: se la distribuzione di frequenza è normale, allora circa i 2/3 delle osservazioni saranno non oltre l'1 di deviazione standard dalla media, e il 95% sarà prima di 2 di deviazione standard.

Abbiamo poi il **COEFFICIENTE DI VARIAZIONE, CV,** che calcola la deviazione standard come **percentuale** della media: **deviazione standard/media campionaria**.

Altra misura molto usata è la MEDIANA, che descrive la posizione di una distribuzione di frequenza, e viene spesso presentata in un BOX PLOT accanto alla distribuzione interquartile.

La **MEDIANA** divide le misure ordinate in due metà, in modo che sia a destra che a sinistra della mediana ci sia esattamente lo stesso numero di misure.

L'**INTERVALLO INTERQUARTILE** è invece la differenza tra il primo e il terzo quartile di dati: il **PRIMO QUARTILE è il 25-esimo percentile, il secondo quartile è la mediana, il terzo quartile è il 75-esimo percentile**.

Misura di **TENDENZA CENTRALE**: dobbiamo ordinare i valori dal più basso al più alto e cerchiamo la misura che sta nel mezzo, misura che tiene sia a sx che a dx un numero uguale di casi. Si sceglie di solito la media aritmetica dei due valori. La misura di tendenza centrale dice che tutti i valori stanno metà da un lato e

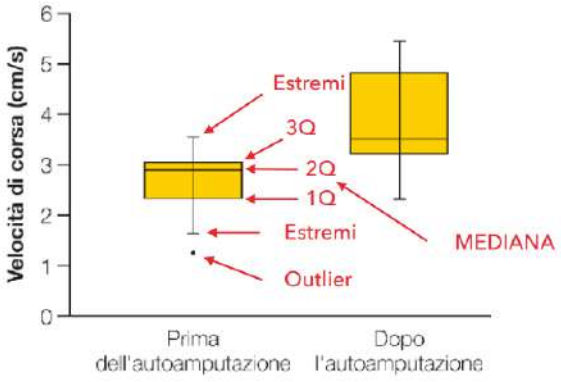
metà dall'altro lato. In questo modo posso calcolare altre misure. La **MEDIANA è la misura che lascia sia a sx che a dx la metà dei casi**. Facendo così **divido la mia serie di dati in QUARTI**: parlo di **QUARTILE**. La mediana è quella a 50/50, mentre il terzo quartile è quello che ha solo il 25% dei casi più grande di lui.
Altro esperimento su mediana e quartile: prima o dopo l'amputazione c'è un aumento della velocità? Valori di velocità, non importa in che unità di misura. Intervallo interquartile: se ci sono tanti casi intorno alla mediana, la distanza interquartile è PICCOLA, perché i casi attorno alla mediana sono più che nelle code, e quindi il 50% centrale sarà vicino alla mediana.

Una MISURA DI TENDENZA CENTRALE è UN SINGOLO VALORE CHE RAPPRESENTA IL PUNTO CENTRALE DI UN SET DI DATI. CI SONO TRE MISURE COMUNI DI TENDENZA CENTRALE: MEDIA, MEDIANA, MODA

BOX PLOT

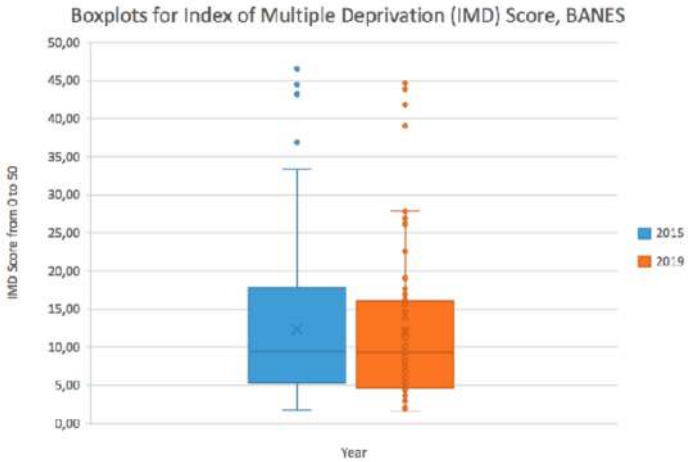
Il BOX PLOT visualizza la mediana e la differenza interquartile. **Ogni scatola (box) di un box plot è delimitato sul lato inferiore dal primo quartile, e sul lato superiore dal terzo quartile. La linea centrale che divide le due box è la MEDIANA**. Le **linee verticali**, i cosiddetti baffi, si estendono verso l'esterno del box e sono **IL PIU' PICCOLO E IL PIU' GRANDE VALORE NON ESTREMO DI DATI. Si indica con valore ESTREMO per definizione un valore che sta distante dal box più di 1,5 volte l'intervallo interquartile**.

GLI OUTLIER NON SI TOLGONO!!!



Il primo estremo è il trattino orizzontale
Il box plot indica quindi DOVE SI TROVA LA MAGGIOR PARTE DELLE OSSERVAZIONI. Un istogramma può farlo in modo più particolareggiato, ma un box plot evidenzia le cose importanti della distribuzione di frequenza.
Come si vede nel grafico sotto, il **PUNTO corrisponde all'UNICO VALORE ESTREMO DI TUTTI QUELLI MISURATI, che appunto vengono indicati con un punto**.

Il puntino esterno è un **OUTLIER**: l'outlier è il caso anomalo, e sta oltre gli estremi. NON SI TOLGONO gli outlier, **SOLO quando il fatto che il valore è un outlier è un errore di battitura, misurazione o è un soggetto anomalo**. Devo prima capire perché è un outlier, e solo dopo lo posso togliere se vedo che davvero è un errore. **Evidenzio l'outlier con la rappresentazione grafica.**



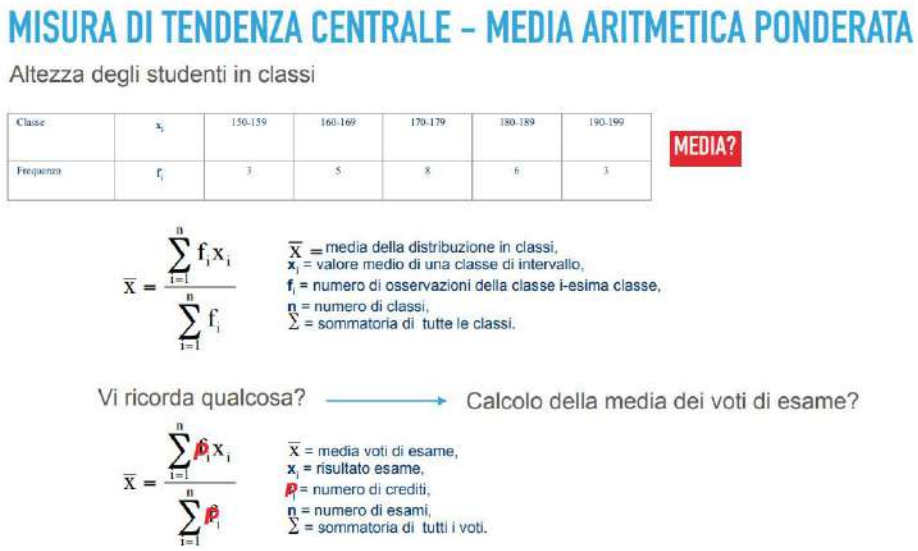
In questa rappresentazione si perde l'APPAIAMENTO. Rappresentazione più corretta del set di dati. Il box plot descrive molto bene la forma della distribuzione, soprattutto se si è abituati a ragionare in termini di quartili.

I box plot si usano quando si vogliono comparare dati numerici tra diversi gruppi.
SIMMETRIA/ASIMMETRIA: quando una distribuzione di dati è simmetrica, ci si aspetta che la mediana sia nel centro esatto

del box, tale per cui la distanza tra Q1 e Q2 sia uguale alla distanza tra Q2 e Q3. Anche i baffi saranno simmetrici rispetto alla mediana, a meno di outliers indicati con il puntino. Se invece la distribuzione non è simmetrica, la mediana non sarà nel centro, e si potranno trovare dei baffi più lunghi e dei baffi più corti, uno con outliers e l'altro no. quindi il box plot riesce a DIRE se un dato è asimmetrico.

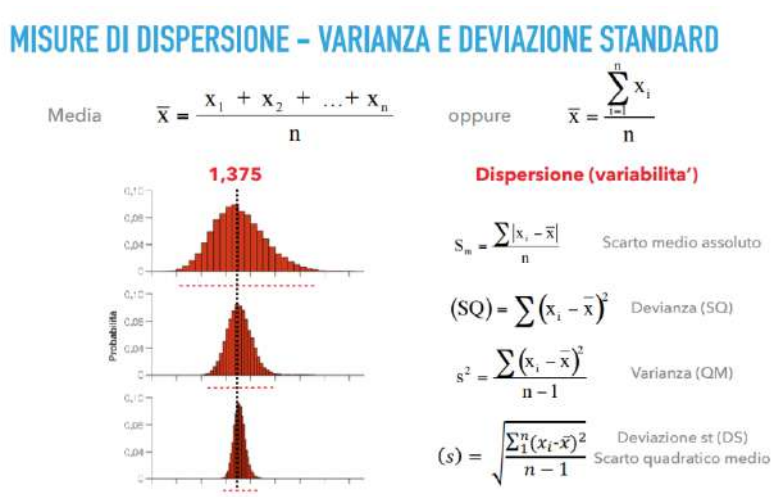
Si può fare sia quello a destra che quello a sinistra

IL NUMERO MINIMO DI DATI PER FARE UN BOX PLOT è 5



Misura di tendenza centrale – **MEDIA ARITMETICA**
Si calcola come sommatoria di tutti i valori misurati diviso il numero totale dei casi. **Media ponderata: prendo quante volte ho un dato valore in una classe**

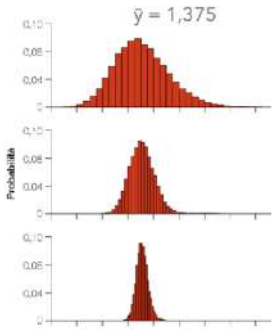
Gaussiana: la MEDIA è il BARICENTRO della distribuzione. Quando una distribuzione è esattamente simmetrica, la MEDIANA e la MODA sono UGUALI. più poi una distribuzione è asimmetrica, più si distanziano. La MEDIA invece è quella che bilancia le distanze a dx e a sx del suo valore, in modo che le distanze sommate diano zero.



Lettere greche: descrizione dei parametri (mu, sigma)
Lettere normali: calcolo la STIMA (differenza tra sigma quadro e S quadro, perché sto STIMANDO con S quadro la variabilità in un campione)
Trucchetto: se N è molto grande e tende all'N campionario, se rimuovo 1 non è un problema. Il fatto di fare N-1 è fondamentale per i campioni bassi, che sono poco rappresentativi. Quindi l'idea è togliere 1, essere più corretto anche se un pochino meno preciso, ma più ROBUSTO. Questo nel caso di una popolazione in cui prendo TUTTI gli individui NON ha senso, perché tanto li ho tutti, cosa che mi dà il valore PARAMETRICO, NON la stima.
Spesso non posso calcolare varianza o deviazione standard, quindi la cosa che faccio è calcolare un **COEFFICIENTE DI VARIAZIONE**, che è una misura molto robusta e intuitiva, e altro non è che sigma diviso mu per la media. Quindi ho S, stima della deviazione standard, diviso la stima della media campionaria!

MISURE DI DISPERSIONE – COEFFICIENTE DI VARIAZIONE

Media $\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$ oppure $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$



Confrontare variabilità?

$CV = \left(\frac{\sigma}{\mu}\right) \cdot 100$

$CV = (s/\bar{y}) \cdot 100$

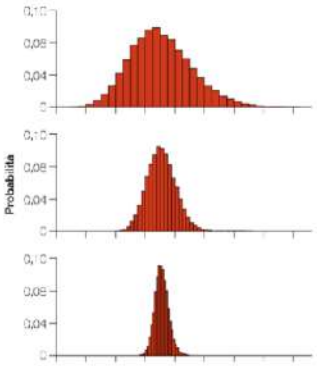
Questo coefficiente di variazione si usa quando voglio confrontare delle misure di variabilità prese in modo diverso; quindi, tipo se la variabilità di un carattere dipende da un altro, tipo peso e altezza!!!

Molto importante calcolarlo se voglio capire se ci sono delle misure che sono legate l'una all'altra

La cosa importante è sapere la deviazione standard, aka scarto quadratico medio, aka ho N campionario scarti quadratici, di cui faccio la media e poi la radice quadrata

PROPRIETA' DELLA MEDIA

$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$ oppure $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$



$\sum (x_i - \bar{x})$ Somma degli scarti = 0

$\sum (x_i - \bar{x})^2$ Devianza (SQ) e' un min

***Associativity*/additiva**

La media delle medie ponderate è = media

Le grandezze additive sono le piu' frequenti in natura

Internalità

Baricentricità

Proprietà della media

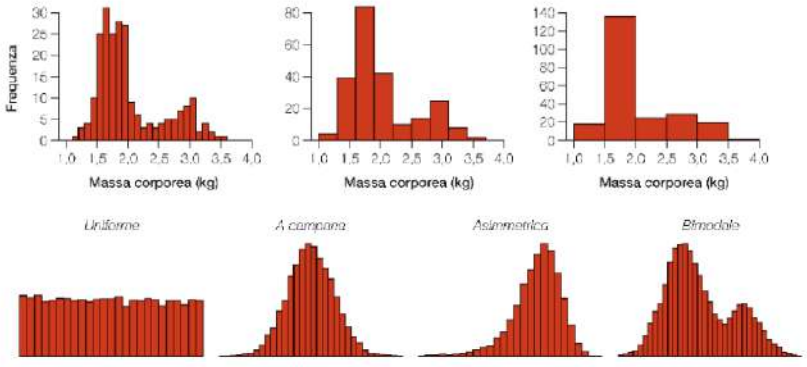
Abbiamo visto per ora **mediana e media come misure di tendenza centrale**.

Non sempre la media è la misura di tendenza centrale più adatta. Per esempio, nel set di dati del salmone rosso, tra le due misure di tendenza centrale scelgo più la **MEDIANA, non la media, perché la media è poco rappresentativa, perché è necessario che il peso delle barre a sx sia uguale al peso delle barre di dx. questo perché ho una distribuzione BIMODALE**. Nel caso in cui avessi una gaussiana o una distribuzione approssimabile ad una gaussiana, andrebbe bene prendere la media, che tra l'altro coincide con la mediana.

DOMANDA – STATS DESCRITTIVE

Salmone rosso (*Oncorhyncu nercka*)

Massa corporea (in gr) di femmine adulte al momento della riproduzione



CONFRONTO TRA MEDIA E MEDIANA

Altro esempio: numero di piastre ossee laterali in genotipi marini di spinarello. Lo spinarello ha una serie di spine che fanno parte del suo apparato difensivo, e che sembrano essere più numerosi dove ci sono più predatori. Sembra che la differenza nel numero di piastre ossee sia principalmente associata ad un unico gene: gli omozigoti dominanti ne hanno tante, gli

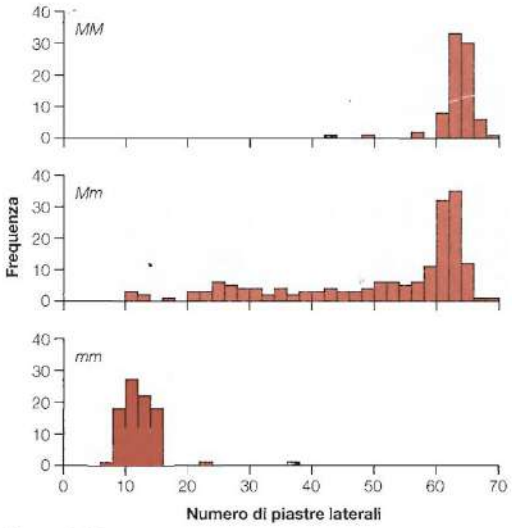


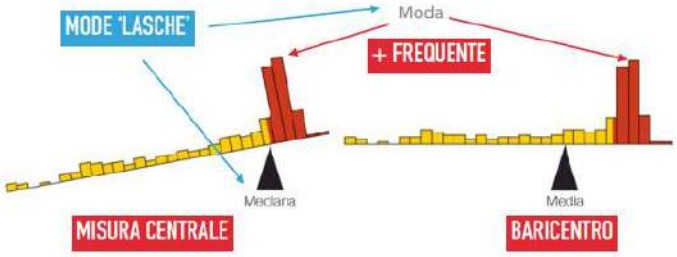
Figura 3.3-1
Distribuzioni di frequenza del numero di piastre laterali in tre genotipi di spinarello, *MM*, *Mm* e *mm*, discendenti da un incrocio tra un nonno marino e un nonno di acqua dolce. Il numero di piastre corrisponde al numero totale di piastre lungo il fianco sinistro e il fianco destro del pesce. Numero totale di pesci: 82 (*MM*), 174 (*Mm*) e 88 (*mm*).

eterozigoti invece un numero molto variabile, compreso in un certo intervallo.
Come si vede nelle immagini, la media è minore della mediana nel caso dei mesci eterozigoti, unico caso in cui si ha una distribuzione fortemente ASIMMETRICA.
SE LA DISTRIBUZIONE è ASIMMETRICA, MEDINA E MEDIA DIFFERISCONO. Questo perché la mediana è l'osservazione centrale di una distribuzione, la media è il baricentro.
In questo caso, la MEDIA è il valore che **si posiziona in modo da BILANCIARE**. Rappresenta il baricentro, mentre le mediana qui tende a stare più verso il PICCO. Nel caso PRECEDENTE, INVECE, ERA LA MEDIANA A STARE NEL MEZZO. **La MEDIANA dà una informazione MENO PRECISA MA ROBUSTA, NON RISENTE DELLA ASIMMETRIA DELLA DISTRIBUZIONE**, motivo per cui è comunque MEGLIO USARE LA MEDIANA, sempre, perché è meno sensibile alle osservazioni estreme. Non ha senso usare la media perché sono più precisa ma molto più sbilanciata.

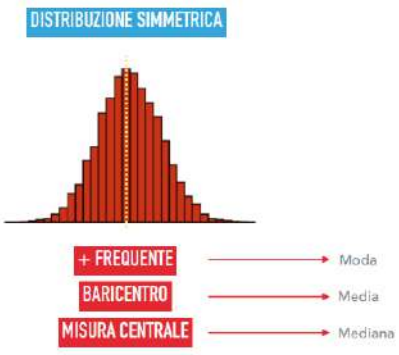
CONFRONTO TRA MISURE DI POSIZIONE E VARIABILITA'

Spinarello (*Gasterosteus aculeatus*)

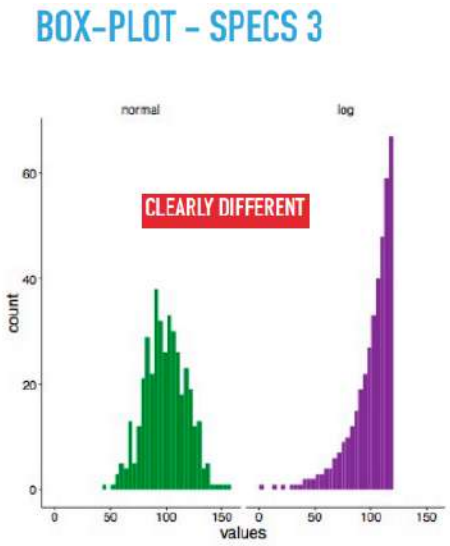
Gene che codifica per ectodisplasina
-> Numero di piastre laterali ossee in genotipi marini *MM*, di acqua dolce *mm* e in eterozigoti *Mm*

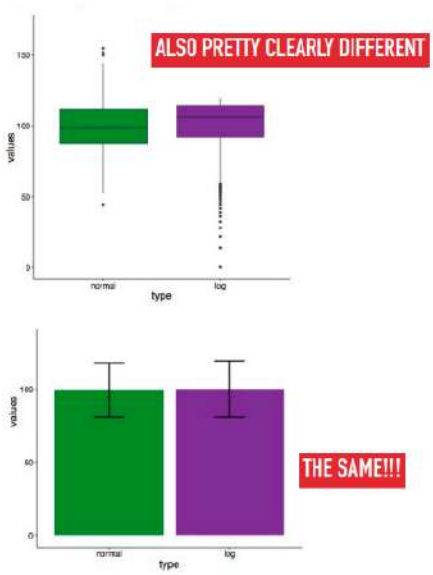


CONFRONTO TRA MISURE DI POSIZIONE E VARIABILITA'



MODA: valore più frequente
La condizione ideale è che moda, media e mediana coincidano. Questo succede nella **GAUSSIANA**





Istogramma e boxplot vanno bene, ma NON I BARPLOTS
Esempio di grafici per far capire quanto è sbagliato usare i grafici stupidi. BISOGNA USARE I PUNTINI!! Il BOX PLOT è il modo miglior di rappresentare le distribuzioni.

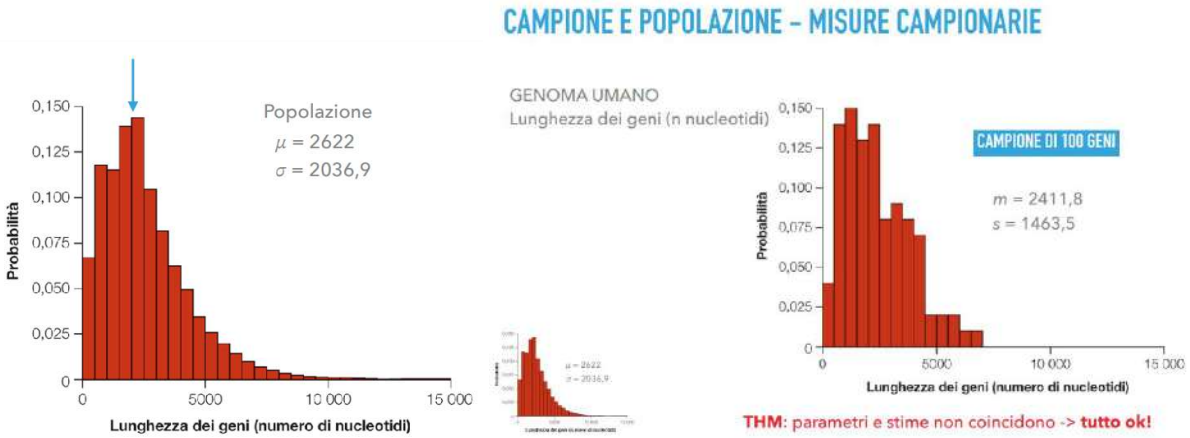
FRIENDS DO NOT LET FIENDS MAKE BARPLOTS

LEZIONE 6

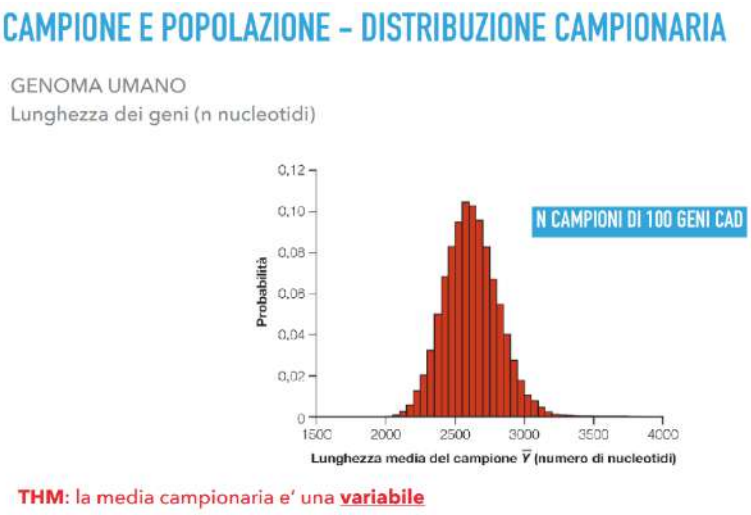
Concetto di **ERRORE CAMPIONARIO**: errore non strumentale o dell'operatore, ma **associato SENZA BIAS al fatto che sto considerando UN SOTTOINSIEME DELLA POPOALZIONE quando campiono**. Quando campiono quindi ottengo una stima diversa dalla media, per il semplice effetto del caso.

DISTRIBUZIONE CAMPIONARIA di una popolazione di cui abbiamo informazioni
La **STIMA** è il processo con cui si inferisce un PARAMETRO della popolazione a partire da dati ottenuti da un campione. La stima è anche il risultato del processo di stima. Sulla base quindi dell'errore campionario, la domanda che ci si fa è quanta fiducia si deve riporre nella stima di fronte al CASO con cui viene scelto un campione. Si usa per questo la **DISTRBUZIONE CAMPIONARIA** della stima, cioè la **distribuzione di PROBABILITA' di tutti i valori di una stima che SI POTREBBERO OTTENERE CAMPIONANDO UNA POPOLAZIONE PIU' VOLTE.**

Per esempio, sappiamo che il genoma è stato sequenziato, e abbiamo l'istogramma di frequenza RELATIVA che descrive la lunghezza dei singoli geni (unità statistica sono i geni) in nucleotidi. Faccio un istogramma e scelgo una larghezza dei BIN, cioè delle colonne.



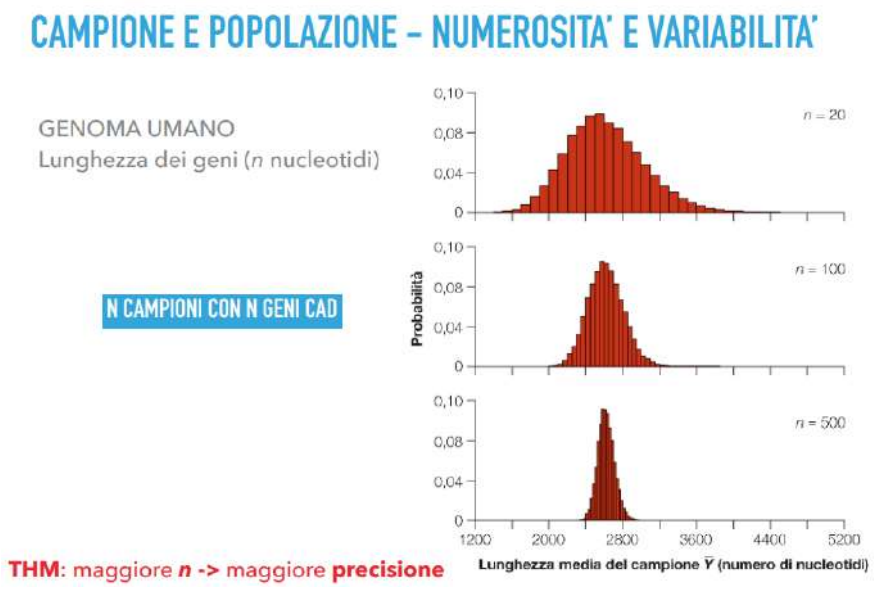
In questo caso, **l'istogramma della figura NON mostra la distribuzione di lunghezza di un campione di geni, ma della POPOLAZIONE DEI GENI**! Per questo, la frazione di geni in un dato intervallo di lunghezze è la probabilità di ottenere un gene con quella lunghezza quando si campiona un singolo gene in modo casuale.
In questo caso, la media e la deviazione standard sono TABULATE, quindi sono espresse in termini di MU e SIGMA, e quindi sono PARAMETRICHE, NON sono una stima perché CONOSCIAMO TUTTA LA POPOLAZIONE.
Ora uso la distribuzione della popolazione come valore di riferimento e cerco di capire cosa succede quando campiono più volte la popolazione nota. Prendo quindi un campione di **100 geni** estratti a caso. Dico al software di estrarre con un generatore di numeri random 100 geni.
Ovviamente la distribuzione di frequenza NON è uguale per il campione di 100 geni, perché sono stati selezionati casualmente. Scrivo un istogramma di frequenza assoluta.
Comunque le due distribuzioni hanno delle cose in comune, come la forma, la dispersione e la localizzazione delle osservazioni più frequenti.
Il m, NON mu, di questa distribuzione, aka la STIMA di mu, sarebbe la **MEDIA CAMPIONARIA** del CAMPIONE, e ha una sua deviazione standard. In questo caso mi viene minore, 60,6%.
Se fosse aumentato, e fosse per esempio andato avanti del 100%, vuol dire che la stima della deviazione standard, che si divide per N-1, mentre la vera deviazione standard è N. **La deviazione standard CAMPIONARIA DI SOLITO TENDE AD ESSERE MAGGIORE DI QUELLA VERA**.



A questo punto posso rappresentare la **distribuzione della MEDIA CAMPIONARIA** della lunghezza dei geni. Cosa vuol dire? Prendo i vari campionamenti di 100 geni, prendo la media di ciascun campionamento e ne faccio la distribuzione (istogramma). Se fossimo in grado di ripetere infinite volte questo processo, otterremmo la distribuzione di probabilità della stima. **Quindi la DISTRIBUZIONE CAMPIONARIA è la distribuzione di probabilità dei valori di una stima che potremmo ottenere campionando una popolazione infinite volte**.

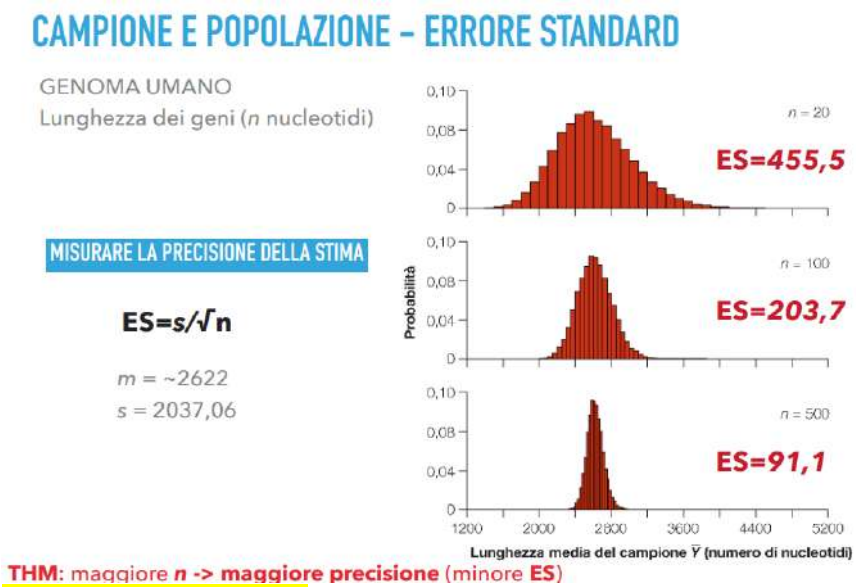
Quindi rappresenta la POPOLAZIONE DI VALORI DI UNA STIMA, non una vera popolazione.

La **distribuzione campionaria è CENTRATA SULLA VERA MEDIA MU. Questo succede se la media campionaria è una STIMA NON DISTORTA (aka non affetta da bias) di MU**.
Ha una forma molto diversa, è una **gaussiana**. Tende a convergere verso il parametro. L'ERRORE CAMPIONARIO dipende dal numero di osservazioni e dalle dimensioni del campione. Sono n campioni di 100 geni. **Se invece di prendere 100 geni ne prendessi 500, la distribuzione sarebbe più stretta, molto molto stretta con un picco altissimo, perché DIMINUISCE L'INCERTEZZA.** Quindi la media campionaria è una variabile diversa dalla lunghezza dei geni. Idem per n uguale a 50, per cui invece AUMENTA L'INCERTEZZA.



Tutte queste distribuzioni della media campionaria HANNO UNA MEDIA (la media della lunghezza media). La media di tutte le distribuzioni, con n pari a 100, 500 e 50, hanno la STESSA MEDIA, la cosa che cambia è l'INCERTEZZA.
Quindi UN CAMPIONE CON N ELEVATO RENDE LA STIMA PIU' PRECISA

Un modo per capire com'è la precisione è calcolare **l'ERRORE STANDARD, che è la DEVIAZIONE STANDARD DELLA DISTRIUZIONE CAMPIONARIA**. Questo errore si calcola NON su una popolazione, perché ho il valore preciso, non ho l'errore campionario. Viene calcolato **OGNI VOLTA CHE SI HA LA STIMA DI UN PARAMETRO**. È dato dalla stima della deviazione standard diviso radice di n. quindi più n aumenta più l'errore diminuisce. **MINORE è L'ERRORE STANDARD, MNORE è L'INCERTEZZA SUL PARAMETRO DELLA POPOLAZIONE.** Quindi la cosa che devo fare è quando calcolo l'errore standard per averlo il più piccolo possibile devo **aumentare il numero di unità campionarie**.

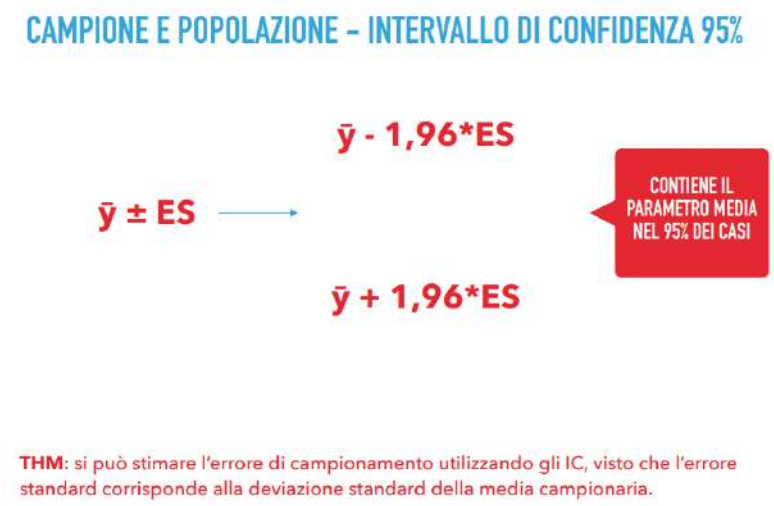


Quindi se ipotizzo media e deviazione standard, come varia l'errore standard nelle tre rappresentazioni della slide (quelle per cui abbiamo supposto 100, 500, 50)? A mano a mano che n aumenta l'errore standard cambia. Maggiore è n, maggiore è la precisione, minore è l'errore standard.

Ogni volta che calcolo una media, ogni volta devo riportare la MISURA DI PRECISIONE ASSOCIATA. Meglio ancora posso fare media -/+ errore standard.

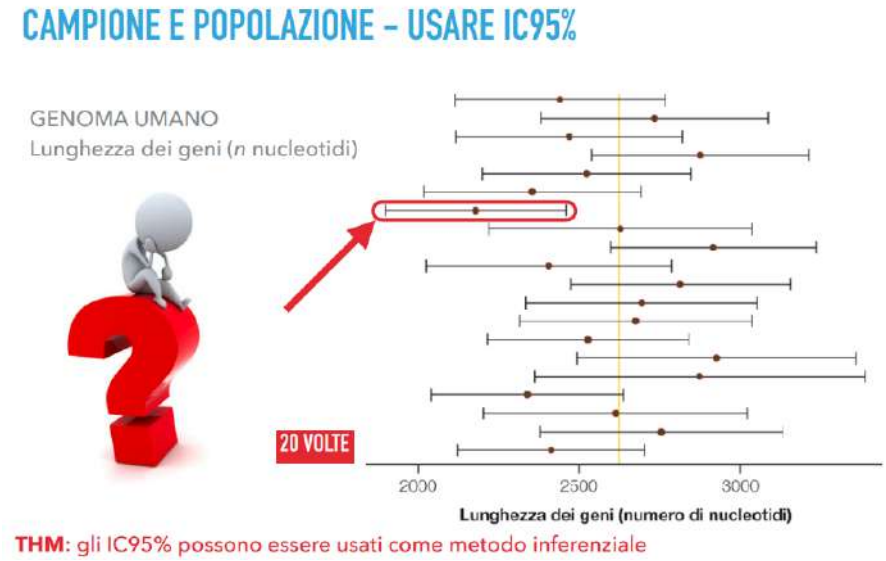
Si può calcolare l'errore standard sia a partire da SIGMA che a partire DIRETTAMENTE DALLA STIMA s, in base al fatto che si voglia l'errore standard della distribuzione parametrica o campionaria

INTERVALLO DI CONFIDENZA



È un intervallo di valori calcolato sulla base dei dati che con una buona probabilità contiene il valore incognito del parametro. Questo intervallo viene calcolato intorno alla stima campionaria. Esempio della media: si prende la media campionaria, e si considera un intervallo a destra e a sinistra della media campionaria in cui si ha una grandissima probabilità di trovare la media parametrica, mu.

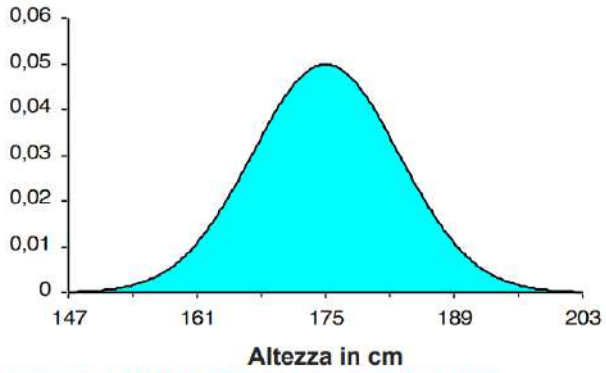
Nel caso della media, l'intervallo di confidenza è al 95%.



Una approssimazione rapida e spesso veritiera dell'intervallo di confidenza al 95% è legata all'errore standard: **SI SOMMA E SI SOTTRAE ALLA MEDIA CAMPIONARIA DUE VOLTE L'ERRORE STANDARD**.

DISTRIBUZIONE 'NORMALE'

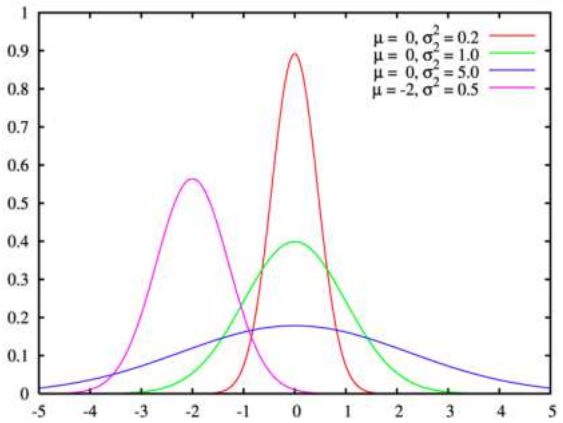
COMUNE PER DATI BIOLOGICI



THM: la curva normale **approssima** molte distribuzioni in natura

DISTRIBUZIONE 'NORMALE'

$$y = f(x) = z = \frac{1}{\sqrt{2\pi \cdot \sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



THM: la curva normale **e' definita da media e deviazione standard**

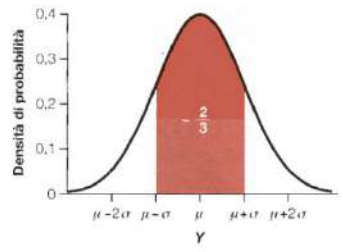


Figura 10.3-1
La probabilità che una misura estratta a caso da una distribuzione normale cada in un intervallo individuato da una deviazione standard dalla media è circa pari a 2/3. (Più precisamente, lo 0,683 delle osservazioni cade entro una deviazione standard dalla media.)

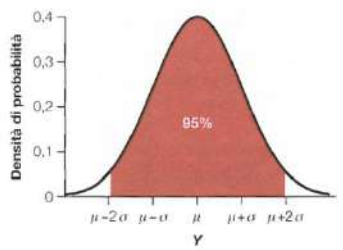


Figura 10.3-2
Esiste una probabilità del 95% che una misura estratta a caso da una distribuzione normale si trovi in un intervallo di circa due deviazioni standard dalla media (per essere più precisi, il 95% giace entro 1,96 deviazioni standard dalla media).

DISTRIBUZIONE NORMALE

È una delle distribuzioni più comuni. Nella slide c'è un istogramma di FREQUENZA RELATIVA in cui i **BIN SONO INFINITESIMI, quindi non è più un istogramma, è una FUNZIONE che rappresenta un istogramma in cui i BIN sono i valori stessi**.

È molto comune. Molte variabili che posso misurare in natura tendono a distribuirsi in maniera NORMALE, in modo simmetrico da un lato e dall'altro seguendo la curva a campana.

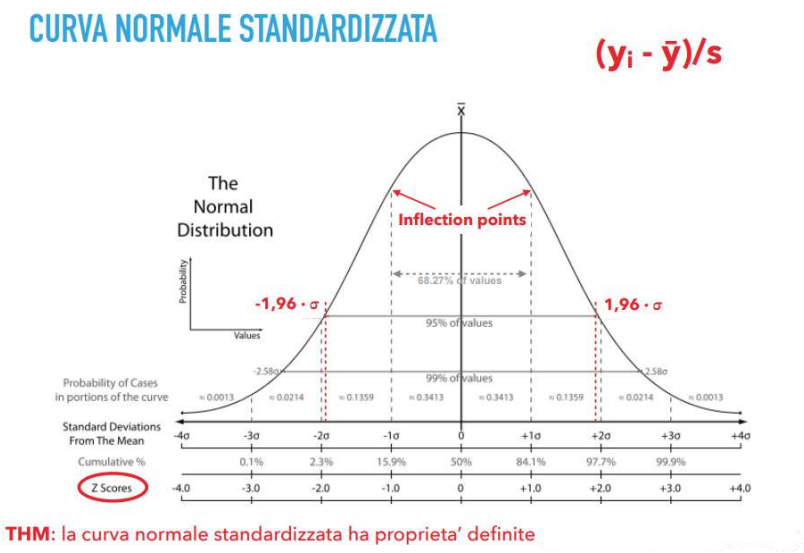
Sono solo DUE i parametri da calcolare per stimare la forma di una curva gaussiana. **È una curva teorica, e necessita di MU e SIGMA QUADRO, come parametri fondamentali**. Le curve normali della slide in cui ci sono **tante curve di altezza e larghezza diversa differiscono tra loro per i loro valori di mu e di sigma quadro**. Si alza o si abbassa anche il picco, perché l'integrale deve sempre essere 1.

Molte variabili in natura approssimano una normale. Per esempio, vediamo la curva di una altezza di un certo numero di soggetti, che NON è una curva teorica ma è precisa, e NON si può confrontare direttamente con la curva della variabile peso. **NON LE POSSO CONSIDERARE E PARAGONARE SE LE VARIABILI SONO DIVERSE!** La soluzione è la **STANDARDIZZAZIONE**: come faccio? Faccio una trasformazione di ogni singolo valore e si rapporta alla deviazione standard. Quindi ci si chiede quanto ogni singolo valore scappa dalla media; quindi, **quanto è lo SCARTO rispetto alla media in unità di deviazione standard**. l'idea è capire quanto un valore dista dalla media. Capire di quante unità di deviazione standard un certo valore è sopra o sotto la media. Questi nuovi valori sono i numeri della nuova gaussiana.

Proprietà della gaussiana:

- Distribuzione continua
- È simmetrica rispetto alla media

- Ha una singola moda
- La densità di probabilità ha il suo valore massima esattamente in corrispondenza della media: in questo modo, media, mediana e moda coincidono.



Caratteristiche: circa i 2/3 (68,3%) dell'area sottesa alla curva corrispondono ad un intervallo tra mu-sigma e mu+sigma
Invece il 95% della probabilità di una distribuzione normale è compreso in un intervallo individuato dal doppio della deviazione standard dalla media (1,96 deviazioni standard): quindi **LA PROBABILITA' CHE LA MISURA DI UNA OSSERVAZIONE ESTRATTA A CASO DA UNA DISTRIBUZIONE NORMALE SIA COMPRESA NELL'INTERVALLO TRA MU-1,96SIGMA E MU+1,96SIGMA è IL 95%.**

La nuova variabile che ottengo, STANDARDIZZATA, avrà **media 0**! La media è infatti sia il baricentro che il valore più frequente, e **se io al valore medio sottraggo il valore medio ottengo 0**! Sta **TRASLANDO e DILATANDO.** La **deviazione standard sarà invece 1**. Uguale alla gaussiana di prima, ma con parametri diversi. Posso farlo con una variabile NOMINALE? NO, perchè NON posso calcolare la media. il valore standardizzato è Z: TRASLATA e DILATATA. Nell'asse verticale c'è la frequenza relativa.

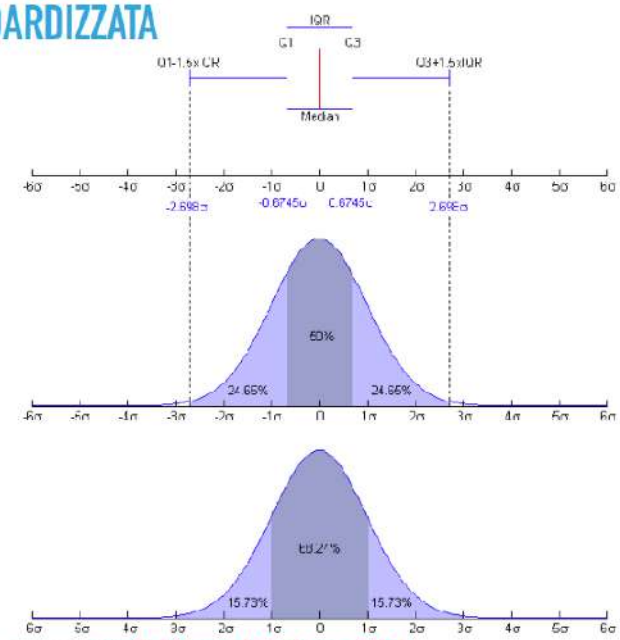
Ovviamente tenendo conto che sigma è 1, l'intervallo in cui si ha il 95% di probabilità di trovare il dato sarà [-1,96,1,96].
Devono essere usate delle TAVOLE, che riportano la probabilità che un valore estratto a caso da una distribuzione normale standard sia maggior di un certo valore critico. Per esempio, **la probabilità che un certo valore estratto a caso sia maggiore di 1,96 è 0,025**! Quindi cerco la probabilità che Z sia maggiore di un dato valore. Si usano sempre solo numeri positivi, tanto la gaussiana è simmetrica rispetto allo zero, quindi i valori negativi e positivi indicano la stessa cosa.

$$Z = \frac{Y - \mu}{\sigma}$$

Questa formula trasforma Y, che ha una distribuzione normale, in un'altra distribuzione normale, che rispetto a Y ha media 0 e varianza 1. Il numeratore indica infatti la distanza di Y dalla sua media, il denominatore la deviazione standard.

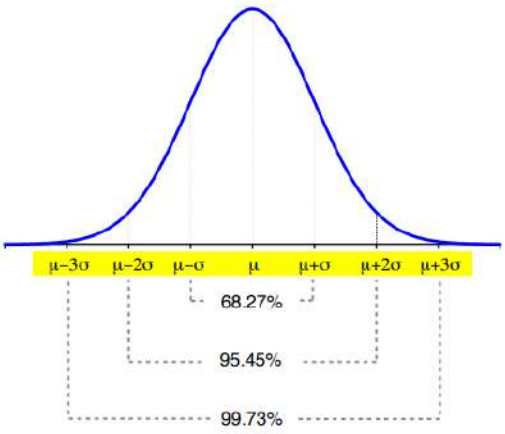
Una distribuzione normale è la DISTRIBUZIONE DELLE MEDIE CAMPIONARIE
Se infatti una variabile ha una distribuzione normale in una popolazione, anche la distribuzione delle medie campionarie avrà una distribuzione normale. La media della distribuzione campionaria è sempre uguale alla media della distribuzione originale mu.
La distribuzione normale standard può essere usata per calcolare la probabilità di estrarre un campione con una media in un determinato intervallo, assumendo di conoscere i valori veri della media e della deviazione standard della popolazione da cui è stato estratto il campione. Tutto questo tenendo conto che si parte da una distribuzione normale delle medie campionarie e si fa una distribuzione normale standardardizzata

CURVA NORMALE STANDARDIZZATA



THM: una distribuzione normale genera **grafici box-plot simmetrici**

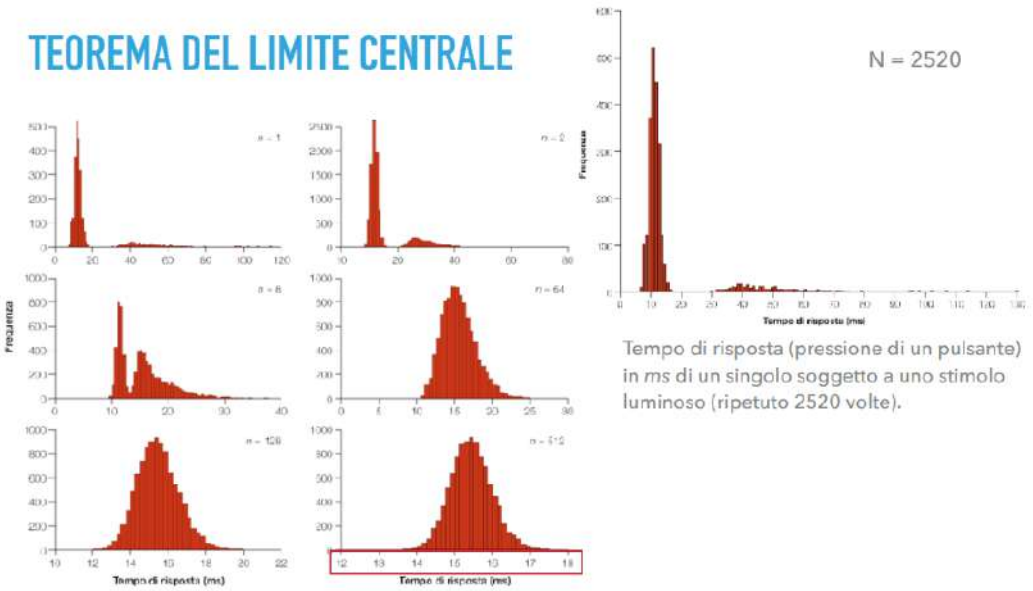
CURVA NORMALE ST. Z



THM: standardizzando i dati si puo' usare la 'normale' come **riferimento per ogni variabile normalmente distribuita**

TEOREMA DEL LIMITE CENTRALE

Le medie dei campioni estratti da una distribuzione normale hanno una distribuzione normale. Inoltre, però, la distribuzione delle medie campionarie ha una distribuzione approssimativamente normale anche se la distribuzione dei dati originali non è normale, a patto che la dimensione del campione sia abbastanza grande. Questo è il TLC.



Come si può vedere dall'immagine, maggiore è il n campionario maggiore è la precisione di approssimazione alla gaussiana. Nel primo caso di prendono infatti campioni casuali di dimensione n=1, e poi si va avanti fino a n=512

TLC: la somma di n variabili indipendenti tutte uguali (con identica distribuzione) è una variabile che si distribuisce normalmente, indipendente dalla forma delle variabili di partenza.

LEZIONE 7

PROBABILITA'

Definizioni: probabilità ed evento

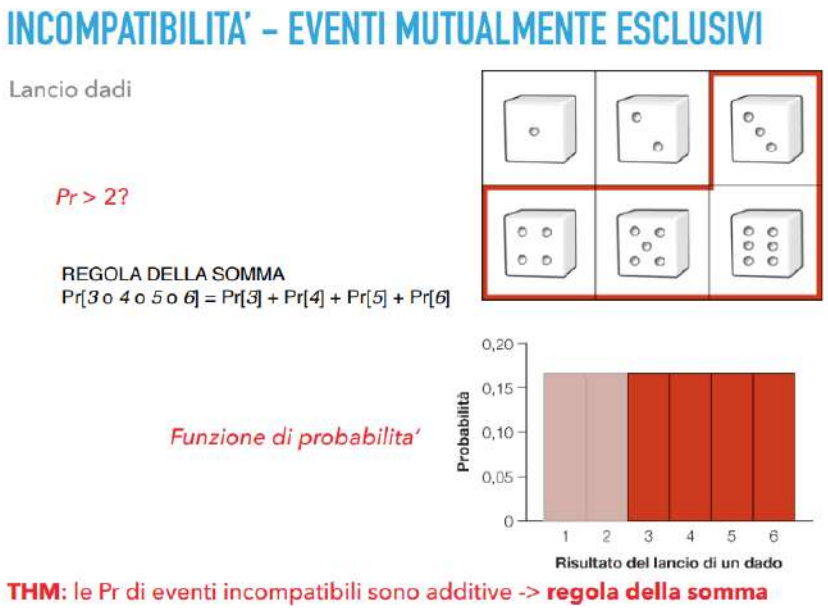
Lo **SPAZIO CAMPIONARIO** è l'insieme di tutti i risultati possibili di una prova casuale, mentre un **EVENTO** è un qualunque sottoinsieme dello spazio campionario. La PROBABILITA' di un evento è la proporzione d tutte le prove casuali con cui si verifica un evento specificato quando la stessa prova casuale viene ripetuta moltissime volte, in modo indipendente e nelle stesse condizioni.

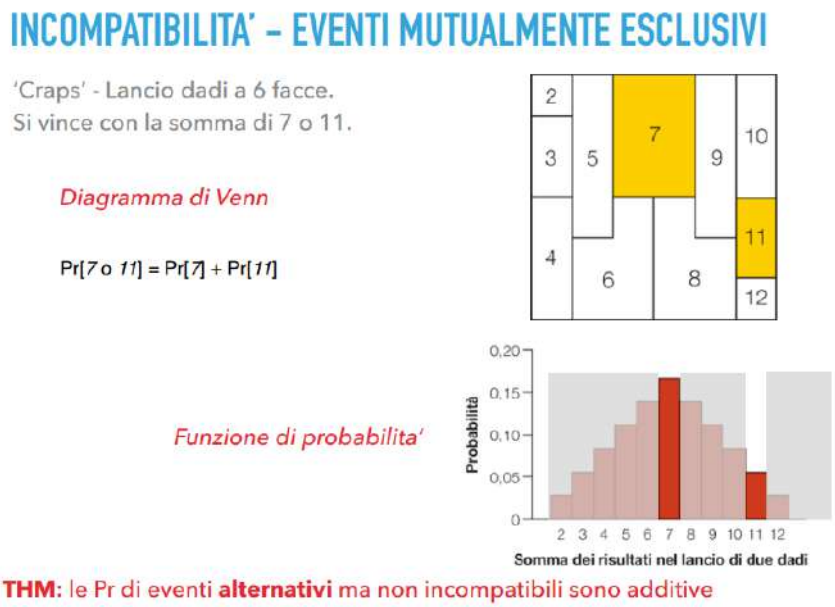
EVENTI INCOMPATIBILI

Due eventi mutuamente esclusivi, quindi che **NON SI POSSONO VERIFICARE CONTEMPORANEAMENTE**, sono INCOMPATIBILI. Quindi la probabilità che si verifichino entrambi è zero.

Le probabilità di **EVENTI INCOMPATIBILI SONO ADDITIVE**: quando due eventi sono incompatibili, la **probabilità che si verifichi uno O l'altro è la SOMMA** delle due probabilità

LA REGOLA DELLA SOMMA RIGUARDA ENUNCIATI CON "OPPURE"





Se poi NON ho eventi incompatibili, la probabilità che si verifichi l'uno o l'altro è data dalla **REGOLA DELLA SOMMA GENERALIZZATA**, che dice che P(AUB)=P(A)+P(B)-P(AeB)

Se gli eventi sono incompatibili, P(AeB)=0, ma non sempre.
Se ho uno spazio di probabilità e prendo tutti gli eventi incompatibili e ne sommo la probabilità, ottengo 1.

INDIPENDENZA E REGOLA DEL PRODOTTO

Due eventi sono **INDIPENDENTI** se il verificarsi di uno non influenza il verificarsi dell'altro, cioè la probabilità che si verifichi il secondo sapendo che è avvenuto il primo è uguale alla probabilità che si verifichi il secondo. Per esempio, se so che è uscito 1 nel lancio dei dadi, la probabilità che esca 2 al secondo lancio sarà sempre 1/6
Da qui si ricava la **regola del prodotto**: la probabilità che si verifichino insieme due eventi indipendenti è il prodotto delle due probabilità
P(AeB)=P(A)P(B)

INDIPENDENZA – ACCADIMENTO DI DUE EVENTI MUTUALMENTE ESCLUSIVI

Lancio dadi.
Accadimento di ottenere 2 volte 3

Diagramma di Venn

Pr[*3* e *3*] = Pr[*3*] x Pr[*3*]

1,1	1,2	1,3	1,4	1,5	1,6
2,1	2,2	2,3	2,4	2,5	2,6
3,1	3,2	3,3	3,4	3,5	3,6
4,1	4,2	4,3	4,4	4,5	4,6
5,1	5,2	5,3	5,4	5,5	5,6
6,1	6,2	6,3	6,4	6,5	6,6

La probabilità di ottenere 3 nel secondo lancio è 1/6

La probabilità di ottenere 3 nel primo lancio è 1/6

THM: la **regola del prodotto** vale per combinazione di eventi indipendenti

EVENTI "SIA CHE" E "OPPURE" – EVENTI INDIPENDENTI

"Sia l'uno o l'altro"

SOMMA
Pr[*A* o *B*] = Pr[*A*] + Pr[*B*]

PRODOTTO
Pr[*A* e *B*] = Pr[*A*] x Pr[*B*]

"l'uno e l'altro"

THM: 'o' -> somma; 'e' -> prodotto

REGOLA PRODOTTO PER PIU' DI DUE EVENTI

Esercizio delle slot machine: un certo tipo di slot machine dà una vincita il 9,8% delle volte che viene usata, e ogni giocata è indipendente dal risultato della giocata precedente. Qual è la probabilità che un giocatore che dispone di monete per otto giocate perda tutto? Per risolvere il quesito bisogna prima calcolare la probabilità di perdere alla singola giocata, che è 90,2%. La probabilità di perdere 8 volte consecutive sarà

quindi il prodotto tra la probabilità di perdere la prima volta E la seconda E la terza… quindi è il prodotto di 90,2% per 8 volte.

ESERCIZIO: SLOT MACHINES

Si vince il 9,8% delle volte indipendentemente dalla giocata precedente
Qual e' la prob che un giocatore perda 8 volte di seguito?

PR[VINCITA]= ?

PR[PERDITA]= ?

Pr[*vincita*]= 0,098

Pr[*perdita*]= 1 - Pr[*vincita*]

PR[PERDITA]= 0,902

Se A e B sono incompatibili
Pr[*A* e *B*] = Pr[*A*]*Pr[*B*]

Pr[*perdita*]= 0,902 * 0,902 * 0,902 … = (Pr[perdita])^8

PR[PERDITA 8 VOLTE DI SEGUITO]= $(0,902)^8 = 0,438$

THM: la **regola del prodotto** vale per combinazione di eventi indipendenti

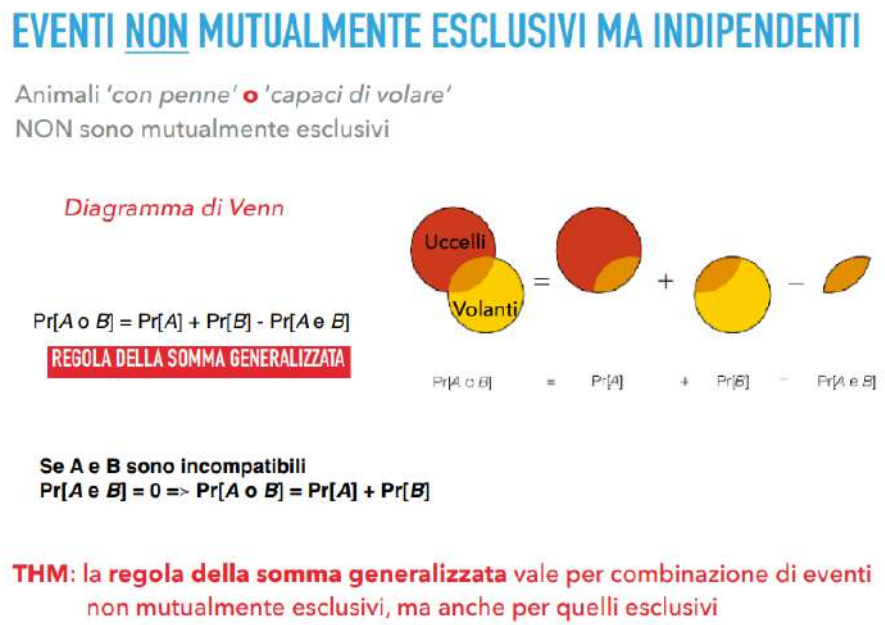
LEZIONE 8

Eventi non mutualmente esclusivi ma indipendenti

Per gli eventi indipendenti la regola della somma e del prodotto ci aiutano a trovare la probabilità degli eventi che avvengono uno O l'altro: somma; l'uno E l'altro: prodotto.

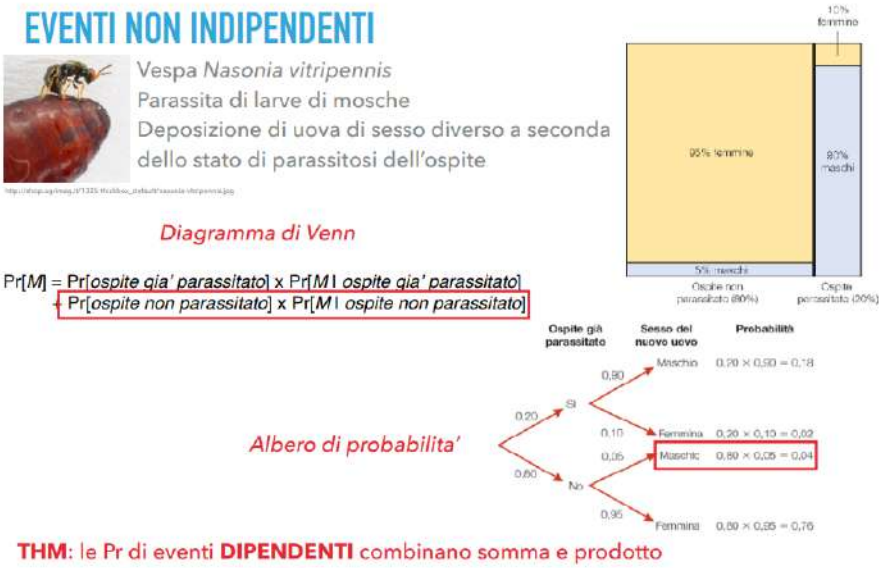
Per eventi NON mutualmente esclusivi si può usare un DIAGRAMMA DI VENN

Regola della somma generalizzata: se gli eventi sono incompatibili gli insiemi sono separati, e quindi il grafico che corrisponde all'intersezione è zero, e dunque di nuovo siamo di fronte alla regola della SOMMA.



Eventi NON indipendenti: probabilità che avvenga una condizione in presenza di un evento che lo condiziona. Vespa che parassita larva di mosche, mettendo le uova nelle larve delle mosche. Probabilità che deponga le uova? Sesso dell'uovo? La vespa è in grado di decidere il sesso dell'uovo. La vespa trova una larva che non ha uova, e può decidere la sex ratio mettendo più femmine possibili e pochi maschi. Se invece è già parassitata, si aspetta di trovare una prevalenza di femmine e una scarsità di maschi, e quindi ne depone lo stesso numero di M e F, così che più o meno la proporzione tra i sessi rimane la stessa. Quando l'ospite è già parassitato si inverte il rapporto, ci sono molti più maschi. **<u>Quindi la probabilità che deponga un uovo maschile o femminile è CONDIZIONATA dalla probabilità che la larva sia già parassitata oppure no</u>**.

Facciamo quindi un ALBERO DI PROBABILITA' per decidere se verranno deposti maschi o femmine.



Per capire come risolvere questo problema, si usa l'albero di probabilità: se la larva era già parassitata, la probabilità che sia maschio o femmina è di un tipo, mentre se non era già parassitata la percentuale cambia. Quali cammini portano al maschio? L'ospite già parassitato e deposizione di uovo maschile, con probabilità P(parassitato)P(maschio), E ospite non parassitato e deposizione di uovo maschile, quindi P(non parassitato)P(maschio). Quindi IN TOTALE:
P(parassitato)P(maschio) + P(non parassitato)P(maschio)

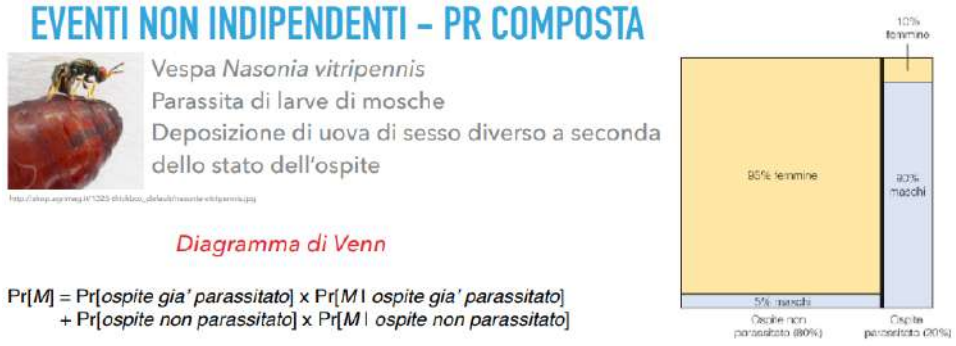
Da qui si ricava il **TEOREMA DELLA PROBABILITA' GLOBALE,** che definisce la probabilità di un evento come la somma delle probabilità che quell'evento di verifichi con tutti i possibili condizionamenti.
PROBABILITA' CONDIZIONATA: probabilità che un evento si verifichi assumendo che se ne sia verificato un altro.
Quindi la probabilità globale di un evento è la somma della probabilità di tutti i valori mutualmente esclusivi (quindi incompatibili) della condizione, associati al proprio singolo condizionamento.

TEOREMA PROBABILITA' GLOBALE $\Pr[X] = \sum(\Pr[Y_i] \times \Pr[X \mid Y_i])$

THM: la Pr condizionata rappresenta la somma di tutti i possibili prodotti (combinazioni)

PROBABILITA' COMPOSTA

Possibilità di trovare la probabilità di due eventi DIPENDENTI usando il condizionamento: la probabilità che si verifichino entrambi gli eventi è la probabilità che si verifichi il primo per la probabilità che si verifichi il secondo sapendo che si è verificato il primo, che è UGUALE alla probabilità che si verifichi il secondo per la probabilità che si verifichi il primo sapendo che si è verificato il secondo (**BAYES**)



TEOREMA PROBABILITA' COMPOSTA
o **regola del prodotto generalizzata**

$$\Pr[A \text{ e } B] = \Pr[A] \times \Pr[B \mid A]$$

Se indipendenti -> Pr[A e B]=Pr[A]*Pr[B], dato che Pr[B | A] = Pr[B]
Ma anche Pr[B e A]=Pr[B]*Pr[A], dato che Pr[A | B] = Pr[A]

Quindi anche $\Pr[A \text{ e } B] = \Pr[B \text{ e } A] = \Pr[B] \times \Pr[A \mid B]$

THM: Il Teorema della ***Prob composta*** funziona anche con eventi non indipendenti ed e' simmetrico

EVENTI NON INDIPENDENTI – TEOREMA DI BAYES

Dato che

$Pr[A \text{ e } B] = Pr[A] \times Pr[B \mid A]$

$Pr[B \text{ e } A] = Pr[B] \times Pr[A \mid B]$

$Pr[A \text{ e } B] = Pr[B \text{ e } A]$

Allora $Pr[A] \times Pr[B \mid A] = Pr[B] \times Pr[A \mid B]$

-> $\mathbf{Pr[A \mid B] = (Pr[A] \times Pr[B \mid A])/Pr[B]}$

TEOREMA di BAYES

THM: Il teorema di Bayes deriva direttamente da quello della prob. composta.

Esempio teorema di Bayes: la probabilità di rilevare un vero positivo di 0,6, mentre di 0,05 di avere un falso positivo. Incidenza della malattia è 1/1000. Voglio così applicare Bayes per trovare la probabilità che sia positivo il test: positivo sapendo che è Down + positivo quando non è Down per la probabilità che non lo sia Qui è molto utile usare l'albero di probabilità.

ESEMPIO/ESERCIZIO SU BAYES

http://www.rivieraweb.it/sites/default/files/styles/large/public/notizie/Sindrome%20di%20Down.jpg?itok=FeKAsylq

Sindrome di Down (DS)
Prevalenza 1/1000 feti
Amniocentesi (1/200 aborti)
Triplo Test (TT) - 0,6 sensitivita' (*pr* di 'true positive')
- 0,05 - *'false positive'* negativo anche quando il TT e' +vo

TEOREMA di BAYES

$Pr[A \mid B] = (Pr[A]*Pr[B \mid A])/Pr[B]$

$\mathbf{Pr[DS \mid TTpos] = (Pr[DS]*Pr[TTpos \mid DS])/Pr[TTpos]}$

$Pr[DS]=1/1000=0,001$

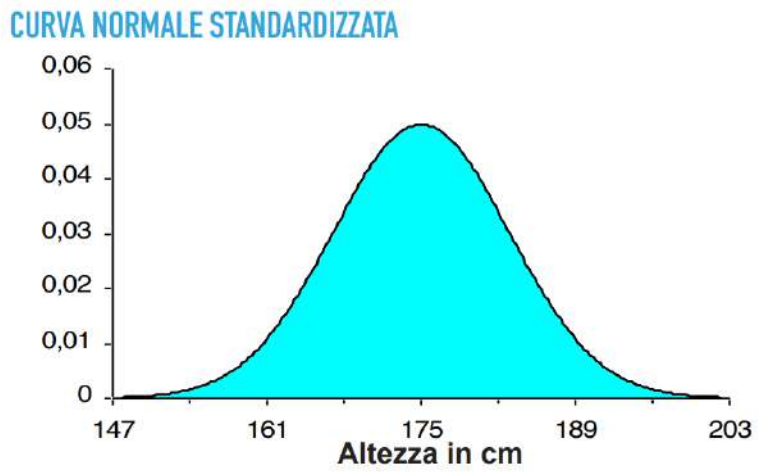
$Pr[TTpos \mid DS]=0,6$

$Pr[TTpos] = Pr[TTpos \mid DS]*Pr[DS] + Pr[TTpos \mid no\ DS]*Pr[no\ DS]$

$Pr[TTpos] = (0,6*0,001) + [0,05*(1-0,001)] = 0,05055$

$\mathbf{Pr[DS \mid TTpos] = (0,6*0,001)/0,05055 = 0,012}$

THM: Il teorema di Bayes sembra complesso, ma non lo e' (troppo)



STATISTICA DI BASE

VERIFICA DELLE IPOTESI

Fondamentale nella ricerca

Curva normale standardizzata. Tabelle che indicano i parametri della gaussiana.

Esempio della distribuzione delle altezze degli individui maschi: curva normale con 175 come picco.

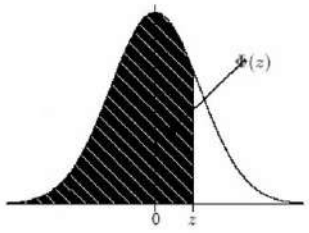
Devo **STANDARDIZZARE LA CURVA**: Z diventa la nuova funzione, data dalla funzione di partenza, MENO il valore medio, il tutto diviso per la DEVIAZIONE STANDARD.

Sapendo che la media dei maschi è **175,6 (MEDIA CAMPIONARIA) e la DEVIAZIONE STANDARD è 7,1, posso calcolare Z.**

Z=(193-175,6)/7,1=2,45

Nella tabella 2,45 corrisponde a 0,9929, il che vuol dire che nella distribuzione normale delle altezze dei maschi il 99,29% dei maschi è PIU' BASSO DI 193 cm.

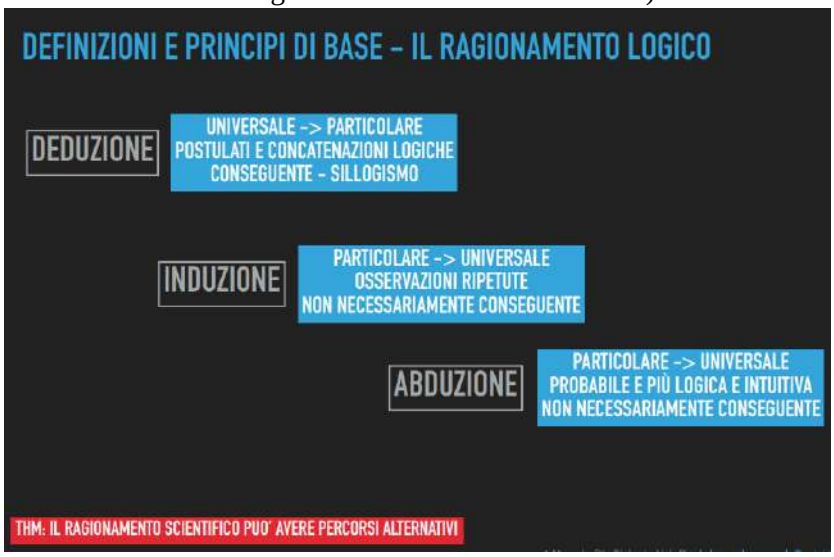
Se cerco il valore con il meno cerco invece la probabilità della coda, dipende da quello che voglio cercare.



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Ragionamento logico

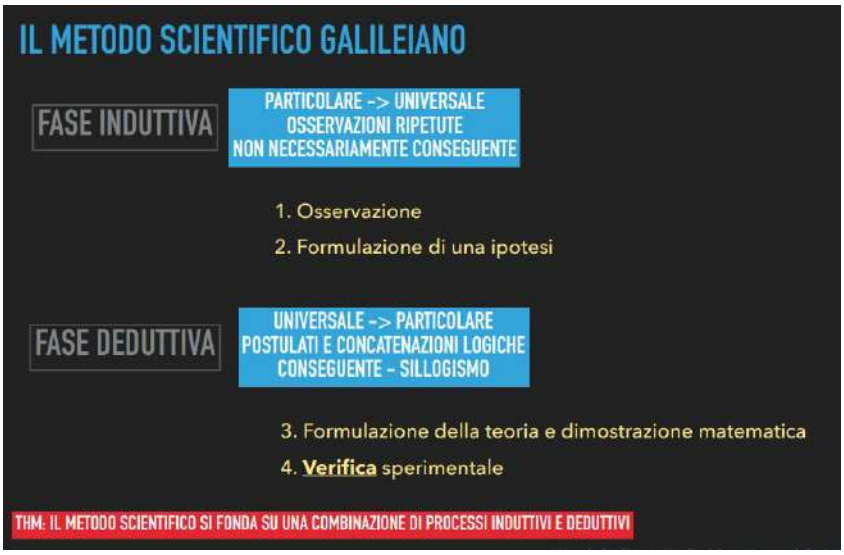
Induzione vs deduzione vs abduzione (da particolare ad universale, probabile e più logica e intuitiva e non necessariamente conseguente. Noi di solito abduciamo).



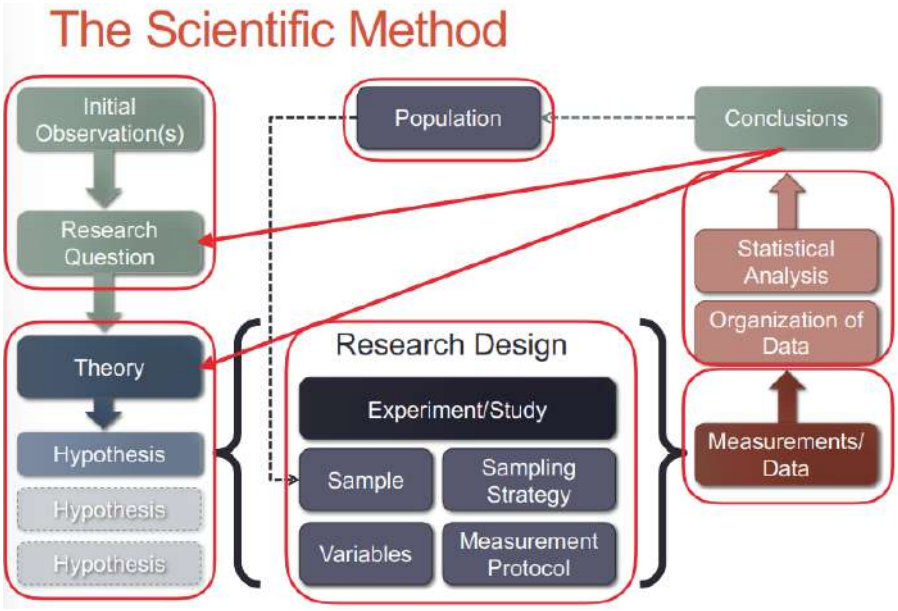
DEDUZIONE: se la regola è che i fagioli sono tutti bianchi, io deduco che i fagioli sono tutti bianchi. Se poi trovo un caso che non vale la regola, vuol dire che o la regola o il caso sono sbagliati, ma NON il processo, perché io parto da una regola

INDUZIONE: prendo dei fagioli dal sacchetto, NON TUTTI, vedo che sono bianchi e penso che siano tutti bianchi. È una regola generale? Non ne sono sicuro! Ma io la impongo perché i casi che ho osservato lo erano tutti

ABDUZIONE: prendo una premessa maggiore certa, una premessa minore dubbia, e una conclusione probabile. L'abduzione è corretta se la regola che abbiamo scelto per spiegare il nostro risultato riceve tante conferme che la probabilità che la regola sia giusta sia ragionevolmente certa. Quindi il mio risultato è che i fagioli sono bianchi, e la REGOLA, che è una cosa che non so, che devo verificare, è che tutti i fagioli del sacchetto siano bianchi. Il caso è che i fagioli che sono bianchi siano del sacchetto, e dato che sono tutti bianchi c'è una conferma alta del fatto che la regola sia giusta, anche se non ne ho la certezza.



Il **METODO SCIENTIFICO** è una combinazione di INDUZIONE e DEDUZIONE.



Definizioni e principi di base

NEOPOSITIVISMO: connubio tra logica e positivismo: criterio di verificabilità: una cosa è vera se può essere verificata. È sufficiente un criterio di verificazione per l'accettazione o validazione di un'ipotesi, teoria o enunciato. Quindi un **enunciato che non possa essere verificato non è necessariamente falso, ma fondamentalmente è privo di senso perché non è sostenuto da prove empiriche**

FALSIFICAZIONISMO: Se non posso confutare una cosa, sto parlando di metafisica. Se una cosa NON si può falsificare NON è scienza.

Popper mio grande amore: il fatto che una teoria non sia falsificabile vuol dire che è metafisica.

Una teoria non falsificabile è la TEORIA DELL'EVOLUZIONE. È talmente tanto affermata che non c'è bisogno di provare a confutarla.

FORMULAZIONE DELLE IPOTESI

Fase fondamentale nel metodo scientifico.

Siamo nella fase **INDUTTIVA**.

Viene poi formulata la teoria, cosa che ci fa iniziare la fase DEDUTTIVA.

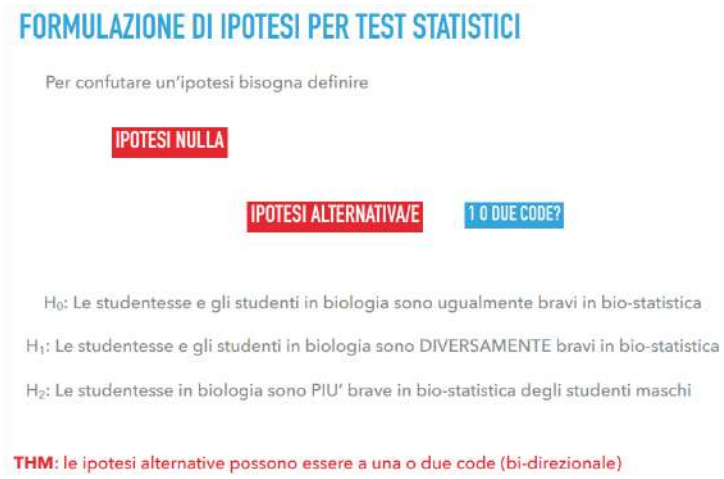
Infine, verifica sperimentale

Formulazione di ipotesi biologica: definisco in primis un **RAZIONALE**, quindi faccio una osservazione, a cui segue un MODELLO, enunciato, regola, poi una PREVISIONE deduttiva, quantitativa, misurabile perché, se non ho questo passaggio la regola rimane non misurabile ed è scritta male. Comunque si definisca il termine ipotesi, ha bisogno di una serie di componenti: **razionale, enunciato e previsioni** precise e misurabili, e DEDUTTIVE.

LEZIONI DA QUI ALLA 13

Per confutare una ipotesi si devono definire: IPOTESI NULLA (H0), IPOTESI ALTERNATIVA (H1, possono essere anche più di una), e scegliere se le **ipotesi alternative siano a UNA O DUE CODE**.

IPOTESI NULLA: è un enunciato specifico riguardo al valore di un parametro della popolazione. Di solito include il punto di vista scettico. Afferma che un parametro della popolazione di interesse è zero (aka NON C'E' CORRELAZIONE).



Di solito è interessante RIFIUTARE l'ipotesi nulla, perché vuol dire che c'è qualcosa di significativo nell'indagine che si sta svolgendo.

L'ipotesi nulla è sempre **SPECIFICA**, quindi identifica sempre un particolare valore del parametro che si sta studiando.

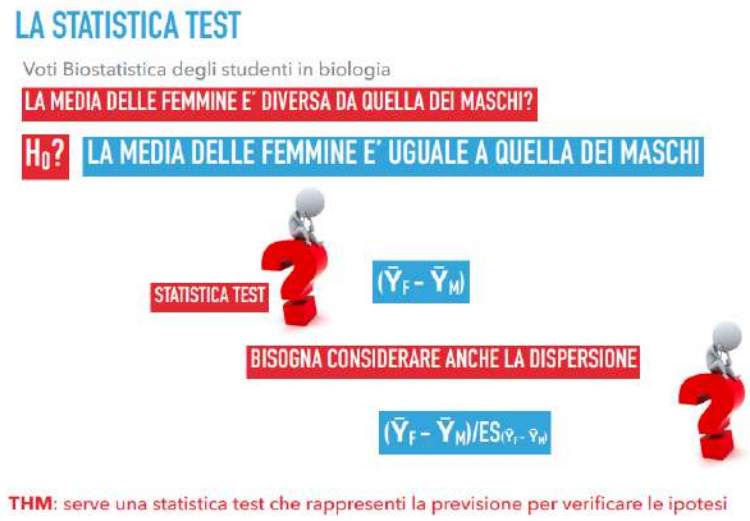
IPOTESI ALTERNATIVA: rappresenta tutti gli altri valori possibili di un parametro rispetto a quelli previsti dall'ipotesi nulla. Quindi spesso coincide con l'ipotesi che il ricercatore vuole sia vera. Per questo motivo, NON è SPECIFICA

L'ipotesi alternativa può essere a una o due code. Se è a DUE CODE, vuol dire che contempla due possibilità: sia il lato MAGGIORE che quello MINORE del valore specificato dall'ipotesi nulla. Si parla anche di TEST BILATERALE, perché vanno bene entrambi i lati della distribuzione di probabilità.

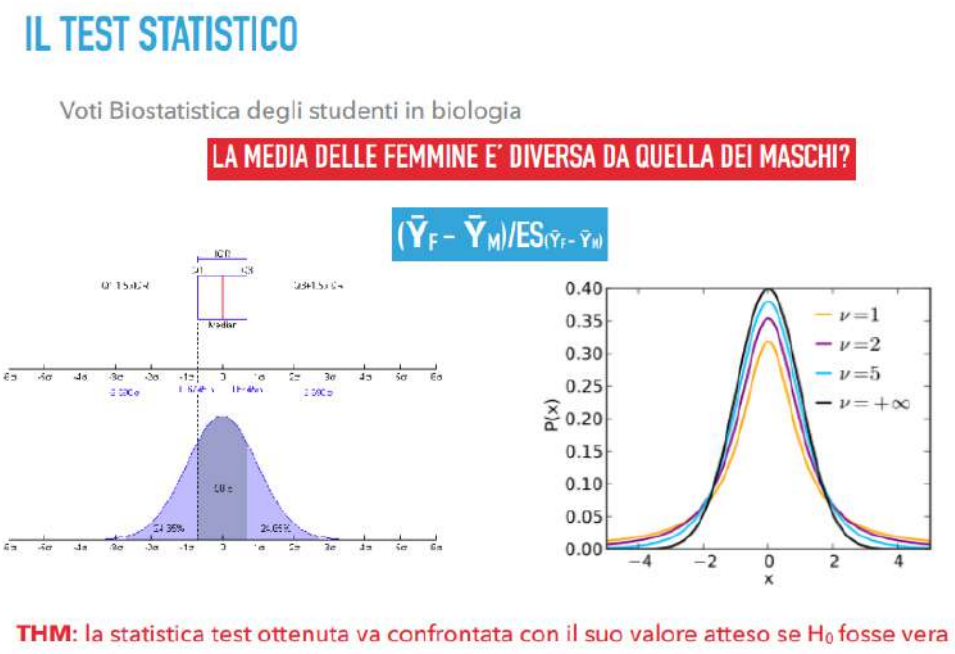
Analizzando i dati, si può vedere se l'ipotesi nulla sia o meno da rifiutare. L'ipotesi nulla, essendo l'UNICA IPOTESI SPECIFICA, è l'unica che può essere rifiutata. Se i dati sono compatibili con l'ipotesi nulla, NON POSSIAMO RIFIUTARLA. **NON SI ACCETTA MAI L'POTESI NULLA**.

LA STATISTICA TEST

È una GRANDEZZA calcolata a partire dai dati per valutare il grado di compatibilità dei risultati con quanto previsto dall'ipotesi nulla.



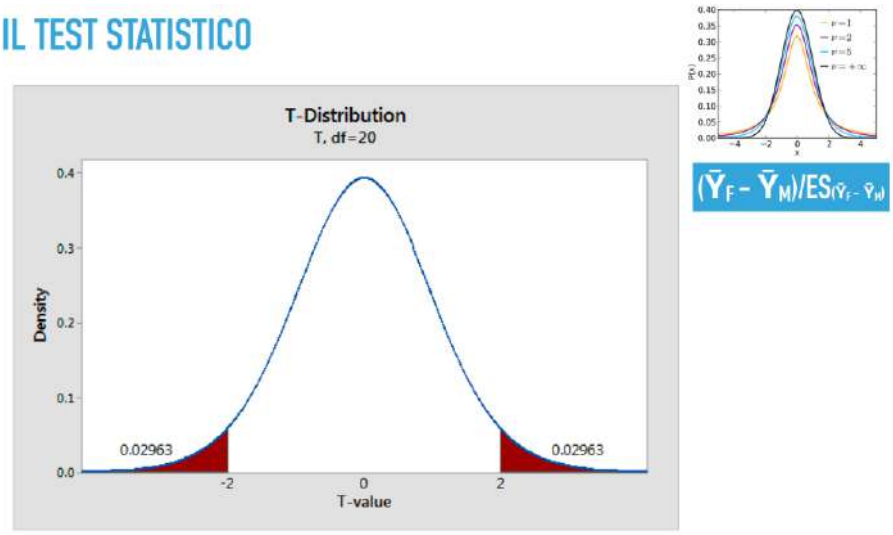
Per esempio, nel caso della media delle femmine e dei maschi, la statistica test è la grandezza che misura la differenza tra la media dei voti femminile e quella maschile. NON BASTA però questo tipo di statistica test. La cosa che viene fatta è considerare anche la DISPERSIONE, cioè la deviazione standard (o l'errore standard). in questo modo si trova la **GRANDEZZA GIUSTA PER VERIFICARE (aka confutare) LE IPOTESI**.



Quindi si può provare a fare un test statistico sulla base della statistica test selezionata. L'ipotesi NULLA è che la media delle femmine sia uguale a quella dei maschi. Si ottiene una distribuzione, che è la rappresentazione grafica della statistica test. **Questa statistica test deve essere confrontata con il valore atteso nel caso in cui l'ipotesi nulla fosse vera**.

Quindi definisco la statistica test, prendo il mio campione e calcolo la grandezza selezionata. **Con quella grandezza stabilisco la funzione di distribuzione campionaria e la confronto con la DISTRIBUZIONE DATA L'IPOTESI NULLA, cioè come verrebbe la statistica test nel caso in cui sia vera l'ipotesi nulla**.
LA DISTRIBUZIONE DI FREQUENZA DELL'IPOTESI NULLA è CENTRATA IN ZERO!!!!!! ovviamente non è zero perché non saranno sempre uguali, ma la media sarà zero obv

La cosa importante da capire è che non sempre i dati rispecchiano la realtà, cioè a causa del campionamento casuale non ci si aspetterà mai di osservare quanto ci si aspetterebbe misurando tutta la popolazione, anche se l'ipotesi nulla è vera. Per decidere quindi se i dati sono o meno compatibili con l'ipotesi nulla (noi di solito speriamo di no), la cosa che si fa è calcolare la probabilità di ottenere una differenza tra la distribuzione campionaria e quella dell'ipotesi nulla grande quando quella che abbiamo ottenuto. Per trovare questa probabilità bisogna determinare come verrebbe la distribuzione campionaria della statistica test nel caso in cui l'ipotesi nulla fosse vera. Questa è la **DISTRIBUZIONE NULLA**.



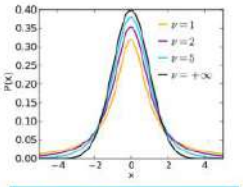
Esempio dei rospi che vediamo dopo: **campionare 18 rospi assumendo che i mancini e i destri siano distribuiti ugualmente nella popolazione è UGUALE A LANCIARE 18 MONETE E VEDERE QUANTE TESTE E CROCI ESCONO**. A questo punto si confronta questa distribuzione di probabilità con quella trovata dal campionamento.

Nel caso media maschi/femmine, la distribuzione NULLA ha come funzione di riferimento quella del test t. Quindi si valutano le tavole del test t per vedere quanto le due distribuzioni sono distanti e se questa distanza sia accettabile rispetto all'ipotesi nulla.

IL TEST STATISTICO

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.600
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
50	0.000	0.679	0.849	1.047	1.299	1.676	2.009	2.403	2.678	3.261	3.496
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										



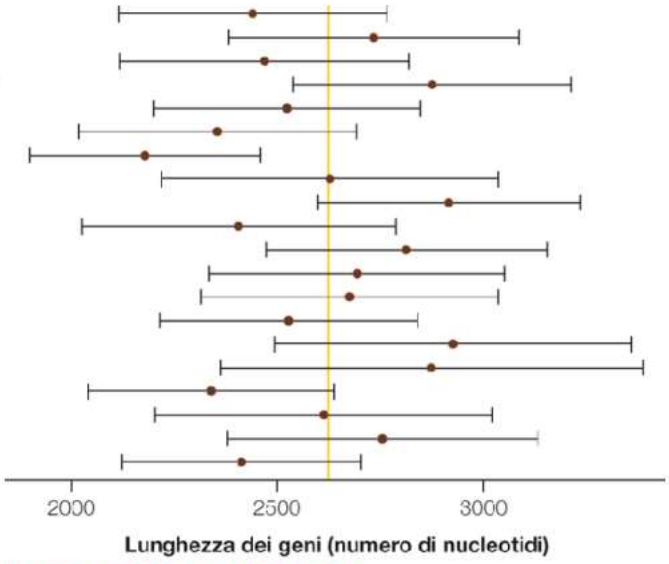
$(\bar{Y}_F - \bar{Y}_M)/ES_{(\bar{Y}_F - \bar{Y}_M)}$

THM: grazie al valore della statistica test si può stimare la probabilità di ottenere tale valore o maggiore per puro caso

DI SOLITO NON SI RIFIUTA L'IPOTESI NULLA SE LA DISTRIBUZIONE CAMPIONARIA è DENTRO IL 95% DELL'INTERVALLO DI CONFIDENZA.

CAMPIONE E POPOLAZIONE – USARE IC95%

GENOMA UMANO
Lunghezza dei geni (*n* nucleotidi)

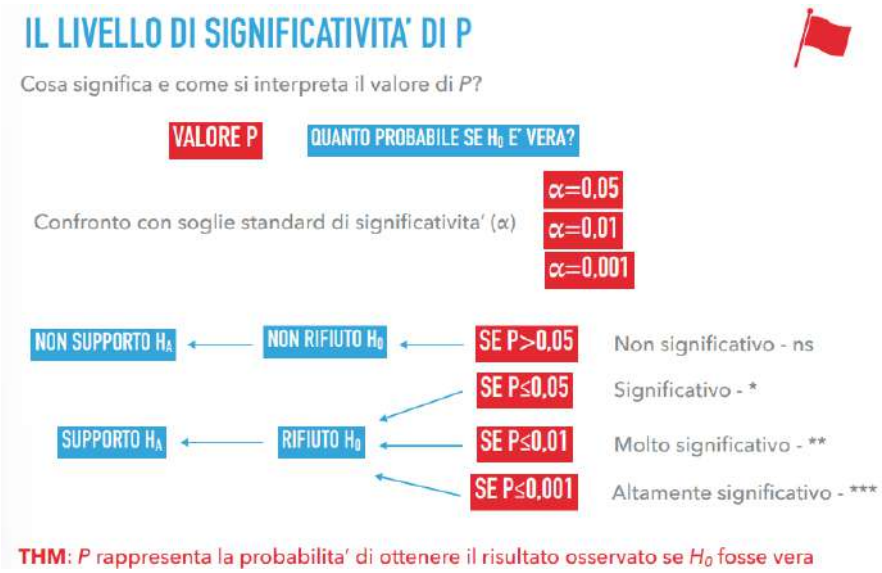


THM: gli IC95% possono essere usati come metodo inferenziale

Manteniamo l'ipotesi dei rospi. Se abbiamo 14 rospi destrimani su 18 non possiamo sovrapporre esattamente la distribuzione nulla con quella campionaria. Ma questo disaccordo è accettabile? È solo dovuto al caso e quindi l'ipotesi nulla è valida? Vogliamo capire quindi se i dati ottenuti da un campionamento casuale sono simili a quelli ottenuti se fosse vera H0, o quelli ottenuti sono troppo insoliti; **vogliamo quindi la probabilità che l'ipotesi nulla sia quella vera**! **Se questa probabilità è piccola allora l'ipotesi nulla sarà considerata incompatibile con i dati e verrà considerata l'ipotesi alternativa. In caso contrario, verrà considerata l'ipotesi nulla**.

La probabilità di ottenere i risultati osservati o i dati meno in accordo con l'ipotesi nulla è detta **p-value**. Mi dice **QUANTO è PROBABILE CHE L'IPOTESI NULLA SIA VERA**. **Se il P value è piccolo, l'ipotesi nulla NON compatibile con i dati e quindi la dobbiamo RIFIUTARE**. Questo P value si calcola in base alla distribuzione nulla della statistica test. Nel caso dei rospi, per esempio, i valori insoliti rispetto alla distribuzione nulla sono in entrambe le code: da 14 a 18 o da 0 a 4 rospi destrimani siamo in un numero troppo insolito rispetto all'ipotesi nulla. In questo caso quindi devo considerare DATI INSOLITI IN ENTRAMBE LE CODE DELLA DISTRIBUZIONE.
P(>=14 destri)=P(14)+P(15)+P(16)+P(17)+P(18)
Questa somma però NON è ancora il p value perché ancora non considera i dati anomali nell'altra coda della distribuzione. In questo caso, visto che è una gaussiana ed è simmetrica, basta MOLTIPLICARE PER DUE. Nel caso dei rospi (vedi sottto) si ottiene 0,031, che è il P VALUE, cioè, dice **qual è la probabilità, assumendo che l'ipotesi sia vera, di un risultato tanto estremo quanto avere uguale o più di 14 rospi destri. Il fatto che il P value consideri anche i rospi da 0 a 4 è perché la distribuzione è simmetrica e mi posso aspettare con la stessa probabilità nella distribuzione nulla 4 e 14 rospi destri. Quindi la probabilità di vedere 14 rospi su 18, o un valore ancora più estremo, è 0,031**.

A questo punto possiamo capire se rifiutare o meno l'ipotesi nulla. Se P è piccolo la dobbiamo rifiutare, se è grande no. **PER CONVENZIONE, P è PICCOLO SE è PIU' PICCOLO DI 0,05**. In questo caso, quindi, rifiutiamo l'ipotesi nulla. Nell'indagine che stiamo facendo, quindi, deduciamo che i rospi destri sono di più di quelli mancini.

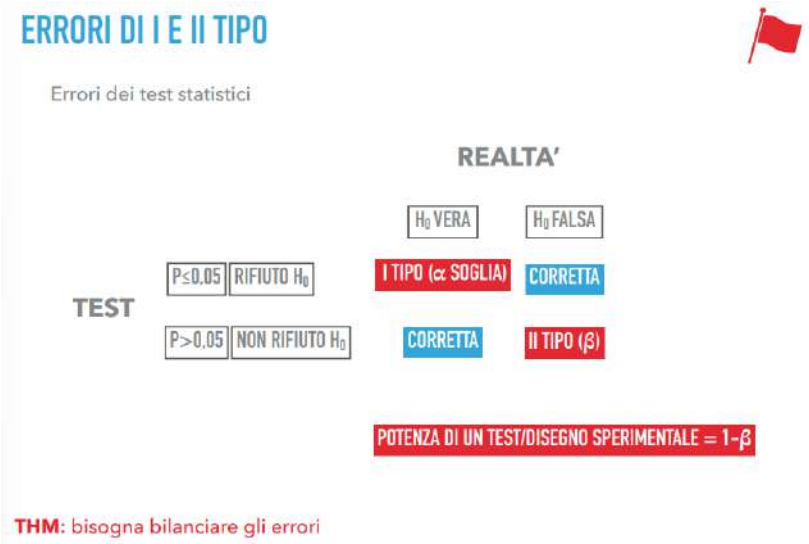


Questa SOGLIA del p value è detta **LIVELLO DI SIGNIFICATIVITA'**, e si indica con **ALPHA**. Di solito, come detto, alpha è 0,05. Il successivo livello usato è 0,01
SE RIFIUTO L'IPOTESI NULLA, SUPPORTO L'IPOTESI ALTERNATIVA!

Quindi come si comunicano i risultati di uno studio? È necessario introdurre: **IL VALORE DELLA STATISTICA TEST, LA DIMENSIONE DEL CAMPIONE E IL P VALUE**. Se presento solo uno di questi è un problema perché non è possibile replicare o accertarsi che i test statistici siano stati fatti nel modo corretto. È anche sempre utile fornire gli **intervalli di confidenza o almeno gli errori standard**.

ERRORI
Se rifiuto l'ipotesi nulla, non è detto al 100% che sia falsa, è sempre una questione probabilistica. Si ha infatti sempre solo l'informazione contenuta nel CAMPIONE, che è influenzata dal caso.
Ovviamente se si è preso un campione casuale non è possibile individuare l'errore nel momento in cui i p value è significativo o non lo è. Però si possono individuare gli errori per capire se ce ne possono essere o no.
Nella VERIFICA DELLE IPOTESI si possono commettere due tipi di errori:

- Errore di **PRIMO TIPO: si rifiuta un'ipotesi nulla vera**
- Errore di **SECONDO TIPO: non si rifiuta un'ipotesi nulla falsa**



In particolare, il livello di significatività di ALPHA (p value) corrisponde alla probabilità di commettere errore di tipo 1: di solito usiamo lo 0,05, cosa che dice che se l'ipotesi nulla fosse vera e noi la rifiutassimo, questa cosa accade una volta ogni 20. **Quindi per ridurre l'errore di tipo 1 è necessario abbassare il valore di alpha. Però questa scelta non è sempre vantaggiosa, visto che aumenta la probabilità di commettere errore di tipo 2**, visto che è molto più difficile rifiutare un'ipotesi tenendo conto del caso con un'alpha così basso. Per questo si usa un compromesso con 0,05.
Al contrario, l'errore di tipo 2 ha a che fare con il mancato rifiuto di una ipotesi nulla non vera. Se in uno studio la probabilità di commettere errore di tipo 2 è bassa, il test ha una elevata potenza.
POTENZA DI UN TEST: **probabilità che un campione casuale determini il rifiuto di una ipotesi nulla falsa**. Non è possibile quantificare la potenza, ma si può dire che **un test ha una potenza maggiore se la dimensione del campione è grande, se la discrepanza dall'ipotesi nulla è grande e se la popolazione è poco variabile**.

Esempio appena fatto dei rospi: ipotesi nulla: i rospi non hanno preferenza nell'uso delle zampe. Ipotesi alternativa: i rospi prediligono una delle due zampe: in questo caso è un TEST A DUE CODE, perché il p value considera sia quando usano poco la zampa sinistra che quando usano poco la zampa destra. Se invece già ipotizzo nell'ipotesi alternativa che viene preferita una sola delle due zampe e scelgo già quale ho un test a UNA CODA.
LE IPOTESI ALTERNATIVE DEVONO ESSERE CHIARE

Come dicevamo, si valuta la distribuzione nulla tenendo conto che l'ipotesi nulla è che non ci sia differenza, e che quindi i rospi sono indifferentemente destri o sinistri. La probabilità diventa quindi il LANCIO DI n MONETE.

Si trova che 14 rospi su 18 usano la zampa destra. Facendo i calcoli si vede che nel test a DUE CODE, tenendo conto del risultato 14, si ottiene 0,031, mentre nel test ad UNA CODA si ottiene 0,0155. In entrambi i casi, comunque, si **RIFIUTA L'IPOTESI NULLA, SUPPORTANDO ENTRAMBE LE IPOTESI ALTERNATIVE**.

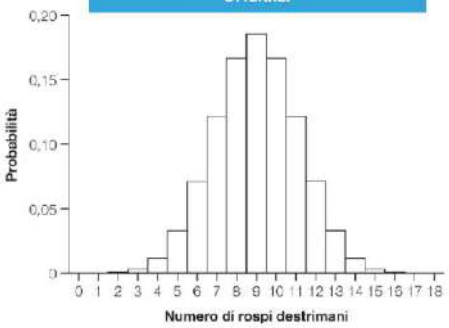
ESEMPIO – ROSPO COMUNE

Rospo comune (*Bufo bufo*)
Togliersi un 'copricapo' con la zampa

18 ROSPI
14 USANO LA ZAMPA DESTRA

SO WHAT?

SE H_0 VERA, P=0,5
FACENDO 18 ESTRAZIONI CASUALI PER N VOLTE
OTTERREI



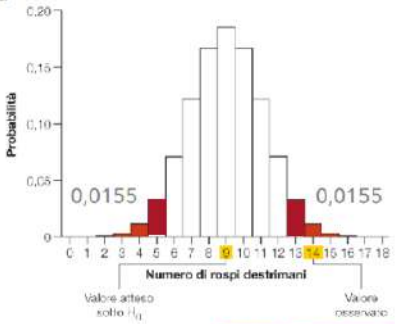
THM: serve un riferimento per stimare quanto probabile osservare per caso (solo come risultato del campionamento) cio' che abbiamo osservato

È interessante tenere conto del fatto che i test unilaterali devono essere usati con cautela, perché il fatto di usare una ipotesi unilaterale rende molto difficile definire anche la distribuzione nulla. Quindi è **CONSIGLIATO USARE I TEST BILATERALI**!

ESEMPIO ROSPI DESTRIMANI

Rospo comune (*Bufo bufo*)
Togliersi un 'copricapo' con la zampa

Se 13/18???



2 CODE P=0,0155+0,0155=0,031
1 CODA P=0,0155

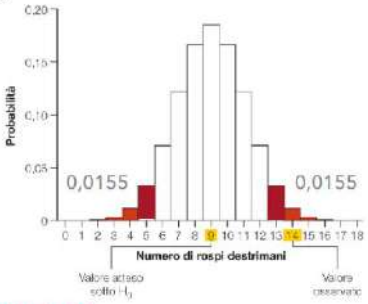
E' poco probabile che H_0 sia vera

RIFIUTO H_0 (SUPPORTO H_A E H_B)
RIFIUTO H_0 (SUPPORTO H_B)

ESEMPIO ROSPI DESTRIMANI

Rospo comune (*Bufo bufo*)
Togliersi un 'copricapo' con la zampa

Se 13/18???



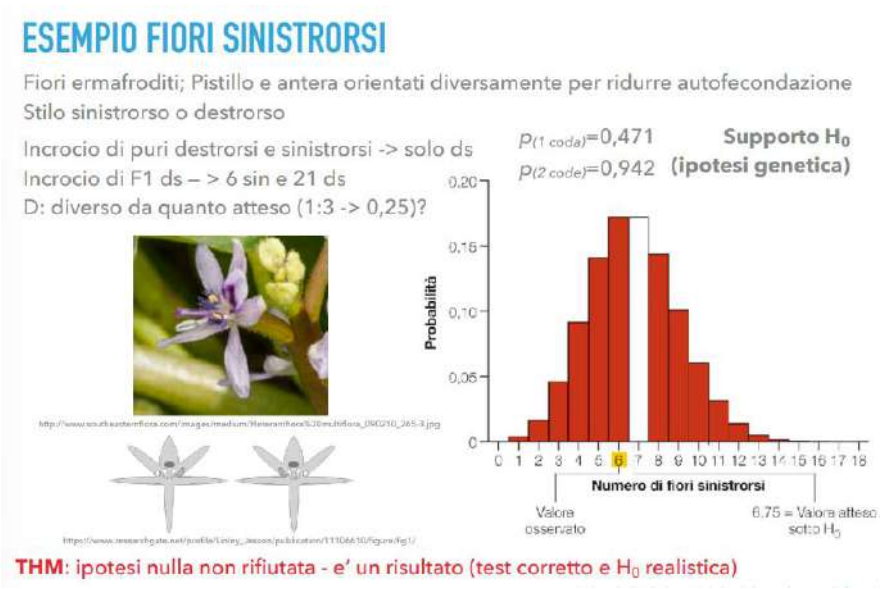
2 CODE P ≈ 0,10 NON RIFIUTO H_0
1 CODA P ≈ 0,05 RIFIUTO H_0

THM: bisogna definire **in anticipo** le ipotesi alternative

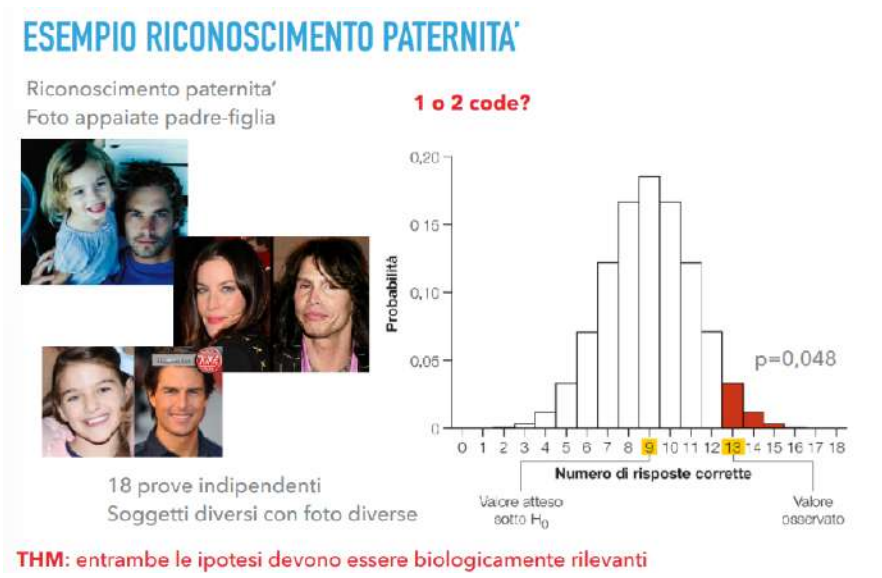
Altro esempio: IPOTESI NULLA NON VIENE RIFIUTATA

Fiori ermafroditi: un meccanismo per evitare che si accoppino con sé stessi è deviare l'organo sessuale femminile verso il lato sinistro in alcuni individui e verso il lato destro in altri individui, con l'antera che si trova nel alto opposto. Per capire come funziona questo tratto geneticamente, sono stati studiati fiori destrorsi e sinistrorsi, incrociandoli e ottenendo solo piante destrorse la generazione successiva. Incrociando poi tra di

loro le piante destrorse, ci si aspettava per la legge della dominanza un rapporto 3:1. Sono state ottenute di 27 piante 6 sinistrorse e 21 destrorse. Questi risultati sono accettabili?
Facciamo un test
IPOTESI NULLA: i sinistrorsi vengono prodotti nella generazione successiva con un rapporto 1:3 rispetto ai destrorsi
IPOTESI ALTERNATIVA: i discendenti destrorsi e sinistrorsi vengono prodotti con un rapporto diverso da 1:3
In questo caso il test è sempre BILATERALE: l'ipotesi alternativa è associata SIA al fatto che i fiori sinistrorsi sono di più che in rapporto 1:3, che siano di meno! NON SI ESCLUDE NESSUNA POSSIBILITA' NELL'IPOTESI ALTERNATIVA.
Il numero di discendenti sinistrorsi è la statistica test, la nostra grandezza, anche se non ha molto importanza se sia sinistra o destra. Assumendo vera l'ipotesi nulla, il numero atteso di sinistrorsi è 27(1/4), cioè 6,75. Questo sarà **MU, il valore medio teorico**.



Poi si **cerca la distribuzione teorica** (sempre usando un computer) per i 27 campioni. Si ottiene una distribuzione discreta simile ad una gaussiana. Sono stati poi osservati 6 fiori sinistrorsi, numero inferiore a 6,75. Si può quindi calcolare il p value. Calcolo:
P(<=6)=P(6)+P(5)+…+P(0)=0,471, che si deve sommare però alla seconda coda, perché il test è bilaterale. Quindi ottengo P(<=6)=2P(di prima)=0,942
Il p value è MOLTO ALTO!
Questo mi dice che l'ipotesi nulla non viene rifiutata. Questo NON vuol dire che L'IPOTESI NULLA è VERA!! La cosa che possiamo dire è che i dati ricavati sono compatibili con l'ipotesi nulla. Ci potrà sempre essere un nuovo studio che permetterà di rifiutare l'ipotesi nulla.

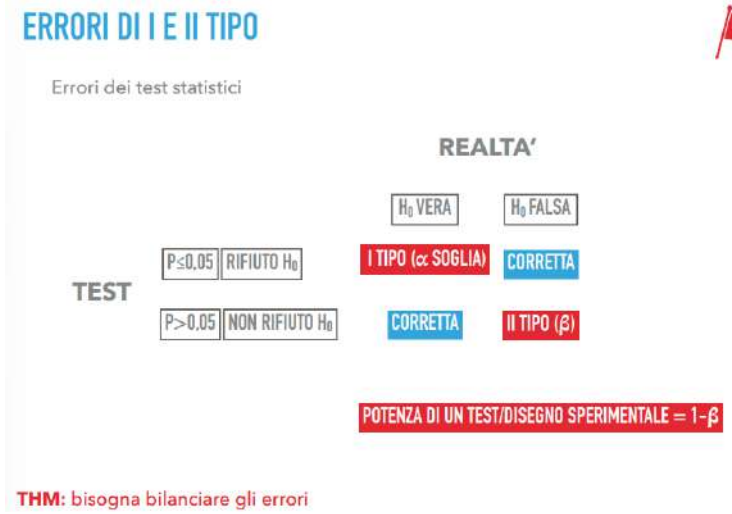


TEST UNILATERALI O BILATERALI?
In un test a **UNA CODA l'ipotesi alternativa tiene conto solo dei valori localizzati ad un lato della distribuzione nulla**. Quindi l'ipotesi viene rifiutata se i dati si discostano dalla distribuzione nel verso specificato dall'ipotesi alternativa.
Un esempio è uno studio che cerca di verificare se le figlie somigliano ai padri. In assenza di somiglianza, la probabilità che il soggetto dia la risposta corretta è 1/2. Quindi si può dire che l'IPOTESI NULLA è che i soggetti scelgono correttamente il padre una volta su due. **L'UNICA IPOTESI**

ALTERNATIVA RAGIONEVOLE è che i soggetti scelgano correttamente il padre PIU' DI UNA VOLTA! Per questo il test è A UNA CODA. **Questo perché i valori dall'altra parte sono concepibili SOLO per il puro effetto del caso.**

Ora immaginiamo di aver condotto 18 investigazioni: da qui, 13 soggetti su 18 hanno indovinato i rapporti di parentela. Assumendo vera l'ipotesi nulla, il valore medio atteso è 9 risposte corrette. Quindi p(>=13)=P(13)+P(14)+…+P(18)=0,0048, che è MINORE DI 0,05. Questo perché NON è necessario dividere per due, perché sto facendo il test ad una coda. **E infatti, se avessimo ottenuto 4 risposte corrette, COMUNQUE AVREMMO DOVUTO CALCOLARE P(>=4) E NON P(<=4)!!!!!**

Quindi ho rifiutato l'ipotesi nulla. Ma a questo punto, qual è l'ipotesi alternativa? È UNILATERALE, cioè che le figlie somigliano più ai padri che agli sconosciuti, NON che le persone NON indovinano a caso la correlazione famigliare.



Supponiamo di voler continuare l'indagine studiando se le figlie scelgono mariti simili ai padri: la ipotesi alternativa è unilaterale? Cioè, i mariti somigliano ai padri, o bilaterale? cioè, i mariti somigliano ai padri o sono molto dissimili ai padri? **IN QUESTO CASO è BILATERALE!! non ha senso non considerare la seconda ipotesi**. DIPENDE DAL CASO, ma è sempre meglio considerarla BILATERALE: **è sempre meglio commettere un errore di tipo 2 che un errore di tipo 1**, perché se faccio un errore di tipo 2 posso sempre andare avanti, mentre se rifiuto una ipotesi nulla vera ho finito.

ANALISI DELLE PROPORZIONI: DISTRIBUZIONE BINOMIALE

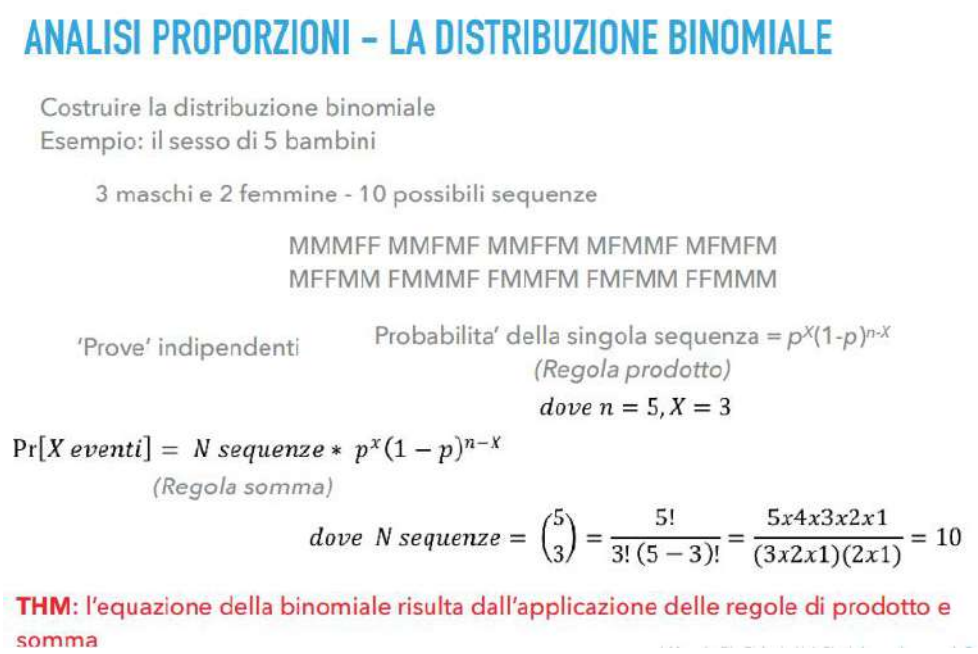
Consideriamo una divisione della popolazione in due gruppi mutuamente esclusivi. Nella popolazione, una proporzione fissa p entra in uno dei due gruppi, gli individui che rimangono fanno parte del secondo gruppo. Se estraiamo un campione casuale di n individui da questa popolazione, la distribuzione campionaria del numero di individui nella prima categoria è descritta dalla distribuzione BINOMIALE, che prende questo nome proprio perché ci sono due risultati possibili, ed entrambi sono rappresentati da categorie con un certo nome.

LA DISTRIBUZIONE BINOMIALE DEFINISCE LA DISTRIBUZIONE DI PROBABILITA' DEL NUMERO DI SUCCESSI IN UN NUMERO FISSO DI PROVE INDIPENDENTI, SE LA PROBABILITA' DI SUCCESSO IN OGNI PROVA è LA STESSA.

La formula dà la probabilità di X successi in n prove, data p la probabilità della singola prova. Si assume quindi che:

- n fisso
- Prove separate indipendenti
- Probabilità di successo p è la stessa in ogni prova

$$\Pr[X \text{ successi}] = \binom{n}{X} p^X (1-p)^{n-X}$$



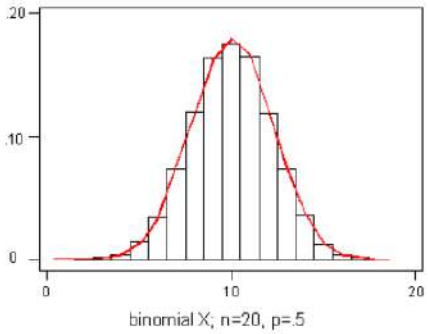
Per esempio, supponiamo di dover campionare il sesso di 5 vespe, dove ogni vespa ha una probabilità di 0,2 di essere maschio. La probabilità che 3 delle vespe siano maschi è (5 su 3)(0,2)^3(1-0,2)^2 **Perché faccio il coefficiente binomiale? Sto semplicemente estraendo 3 maschi da un gruppo di 5 vespe, senza ordine**.

Sotto la descrizione grafica della distribuzione binomiale

ANALISI PROPORZIONI – LA DISTRIBUZIONE BINOMIALE

$$\Pr[X\ eventi] = \binom{n}{X} p^X (1-p)^{n-X}$$

$$dove\ \binom{n}{X} = \frac{n!}{X!\,(n-X)!}$$



THM: la distribuzione binomiale e' la curva di riferimento per stimare la Pr di osservare i nostri dati binomiali se H_0 e' vera

ESEMPIO FIORI SINISTRORSI

Pr[6 sin, 21 des] se H_0 fosse vera?

H_0:
Pr sinistrorsi 0,25 (1:3)
Pr destrorsi 0,75

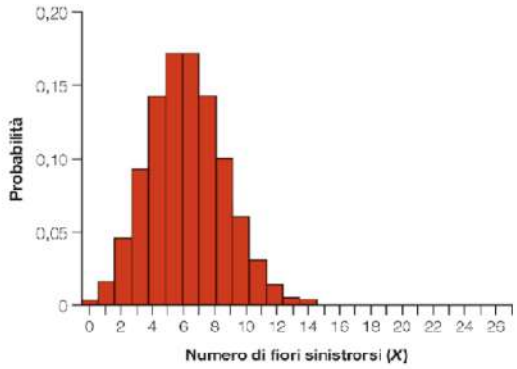
$$\Pr[6\ sinistrorsi] = 0{,}1719$$

se H_0 fosse vera!

$$\Pr[X\ eventi] = \binom{n}{X} p^X (1-p)^{n-X}$$

$$\Pr[6\ sinistrorsi] = \binom{27}{6} 0{,}25^6 (1-0{,}25)^{27-6}$$

$$dove\ 6\ sinistrorsi = \binom{27}{6} = \frac{27!}{6!\,(27-6)!} = \frac{27!}{6!\,21!} = 296.010$$



THM: il test binomiale e' efficiente

Esempio dei fiori sinistrorsi
Stessa situazione di prima: 1/4 di fiori sinistrorsi e 3/4 di fiori destrorsi. Di nuovo si fa una estrazione di un campione casuale da una popolazione di 27 fiori. Il procedimento è proprio associato alla distribuzione BINOMIALE! Ho infatti n prove casuali indipendenti, in cui la probabilità di successo di ogni prova è p, cioè 1/4, che è la probabilità teorica di avere un fiore sinistrorso.
Quindi uso esattamente la distribuzione binomiale! Così posso trovare tutte le probabilità della distribuzione campionaria. **La distribuzione che si ottiene è simile a quella che avevamo ottenuto prima, solo che è ESATTA, perché è stata individuata la funzione corretta**.

DISTRIBUZIONE CAMPIONARIA DI UNA POPOLAZIONE: dati X successi in n prove, la PROPORZIONE DI SUCCESSI è "p HAT", o proporzione campionaria, cioè il numero di successi sul numero totale di prove. Possiamo quindi rappresentare la distribuzione campionaria della popolazione, in cui però la variabile non è più X, ma p hat, cioè X diviso per n: **la MEDIA della distrbuzione CAMPIONARIA di p hat è p**!
Ora, la cosa che cambia rispetto alla distribuzione di probabilità di X è che la distribuzione campionaria di X/n è tanto meno dispersa quanto più alto è n, e questo sulla base di come si calcola la deviazione standard. quindi **l'ERRORE STANDARD DECRESCE ALL'AUMENTARE DELLA DIMENSIONE DEL CAMPIONE**.

ESEMPIO FIORI SINISTRORSI

$$Pr[X\ eventi] = \binom{n}{X} p^x (1-p)^{n-x}$$

H_0:
Pr sinistrorsi 0,25 (1:3)
Pr destrorsi 0,75

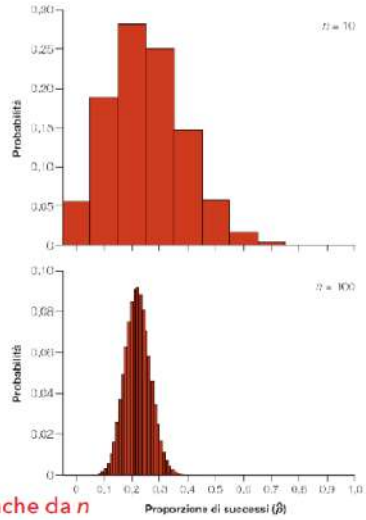
AUMENTO N DIMINUISCO VARIABILITÀ

Distribuzione campionaria di p

$$\hat{p} = \frac{X}{n}$$

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \qquad ES_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}$$

$P_{(60;270)} = 0.033$



THM: il valore binomiale dipende da p ma anche da n

TEST BINOMIALE: esempio con i geni di mus musculus. Il **TEST BINOMIALE usa la distribuzione binomiale per fare una VERIFICA DI IPOTESI**. Si vuole verificare se la frequenza relativa di successi nella popolazione è pari ad un valore atteso assumendo vera l'ipotesi nulla.
IPOTESI NULLA: la frequenza dei successi della popolazione è p0
IPOTESI ALTERNATIVA: la frequenza relativa dei successi della popolazione è diversa da p0.

Esempio: 25 geni di topo coinvolti nella spermatogenesi, di cui è stata individuata la posizione sul genoma. Di questi, 10 sono sul cromosoma X. Se i geni della spermatogenesi fossero distribuiti casualmente in tutto il genoma ci attenderemmo che solo il 6,1% sia sull'X, dato che il cromosoma X ha il 6,1% del totale dei geni di topo.

ESEMPIO GENI MUS MUSCULUS

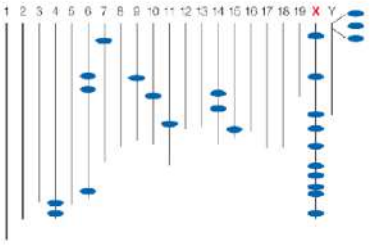
25 geni spermatogenesi
Cromosoma X contiene il 6,1% dei geni
H_0: I geni per la spermatogenesi non sono posizionati preferenzialmente sul X
Pr gene per cromosoma = ?
H_a: I geni per la spermatogenesi sono posizionati su X in maniera non proporzionale alla loro distribuzione sugli altri cromosomi $Pr[genes\ cromosoma\ X] \neq 0{,}061$

10 GENI SU X

Pr[10 crom X, 15 other] se H_0 fosse vera?

$$Pr[X\ eventi] = \binom{n}{X} p^x (1-p)^{n-x}$$

$$Pr[10\ geni\ su\ X] = \binom{25}{10} PrGeneX^{10}(1-PrGeneX)^{25-10}$$



RIFIUTO H_0 **CONSIDERO H_A**

2*Pr[almeno *10 geni su X*] = 1,98 x 10^{-6} se H_0 fosse vera!

$$\hat{p} = \frac{10}{25} = 0{,}40$$

THM: il test binomiale e' efficiente

IPOTESI NULLA: la probabilità di "successo" (cioè, di avere il gene sul cromosoma X) è 6,1%
IPOTESI ALTERNATIVA: la **probabilità di successo NON è 0,061**, cioè i geni possono essere presenti sul cromosoma X più o meno frequentemente (aka la probabilità di successo è diversa da 0,061).
In questo caso **l'ipotesi nulla è molto specifica**, l'ipotesi alternativa no: **TEST A DUE CODE/BILATERALE**. Questo perché potrebbero benissimo esserci meno del 6,1% dei geni della spermatogenesi sul cromosoma X.

Ora dobbiamo identificare la STATISTICA TEST che usiamo per confrontare il risultato osservato con quello atteso. NEL CASO DEL TEST BINOMIALE, LA STATISTICA TEST è SEMPRE IL NUMERO DI SUCCESSI OSSERVATO. Ci attendiamo in media che la frazione del 6,1% dei geni per la spermatogenesi, cioè 1,525, sia localizzata sul cromosoma X. Quindi il **numero di geni NEL CAMPIONE sul cromosoma X è MAGGIORE DI QUANTO ATTESO**.

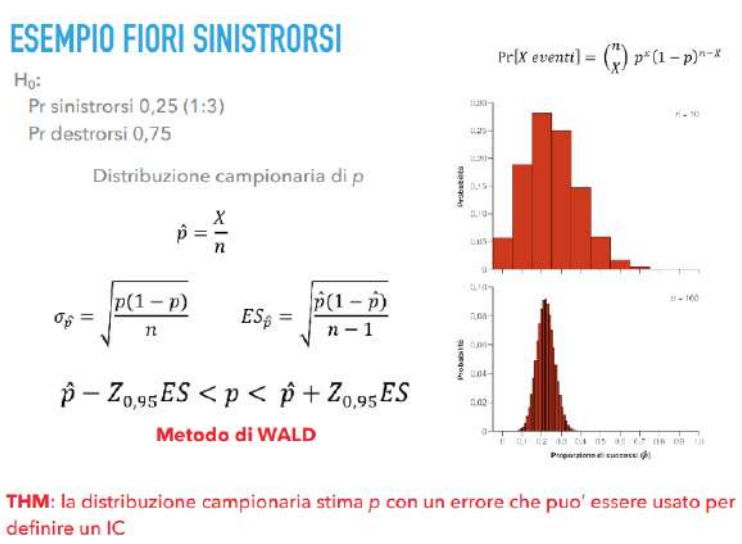
Quindi ora ci si deve chiedere quanto sia improbabile ottenere un eccesso di geni per la spermatogenesi sull'X (aka rifiuto l'ipotesi nulla?)

Dobbiamo quindi capire **COME è FATTA LA DISTRIBUZIONE NULLA**, cioè la distribuzione della statistica test assumendo che l'ipotesi nulla sia vera. **Questa distribuzione nulla è la BINOMIALE: numero di successi X in un campione casuale di n individui estratto da una popolazione con proporzione p=6,1% di successi**. Quindi se ho 25 geni, la distribuzione di probabilità di avere X successi sarà (25 su X)(0,061)^X(1-0,061)^(25-X)

Grazie alla distribuzione nulla posso calcolare il p value, cioè la probabilità che il risultato ottenuto sia estremo uguale o più di 10 geni sul cromosoma X, quindi P=2P(X>=10)

Sapendo che la probabilità di avere esattamente 10 successi è (25 su 10)(0,061)^10(1-0,061)^15, a cui si deve sommare la probabilità di avere un numero più alto di 10 geni, moltiplicando per 2 perché ho un test a due code.
Ottengo come risultato finale 1,98x10^-6, che è MENO di 0,05. Quindi **RIFIUTO L'IPOTESI NULLA**.
Posso anche fare una ulteriore verifica cercando la stima della proporzione, cioè 10/25, che è 0,4, ed è di gran lunga superiore a 0,061

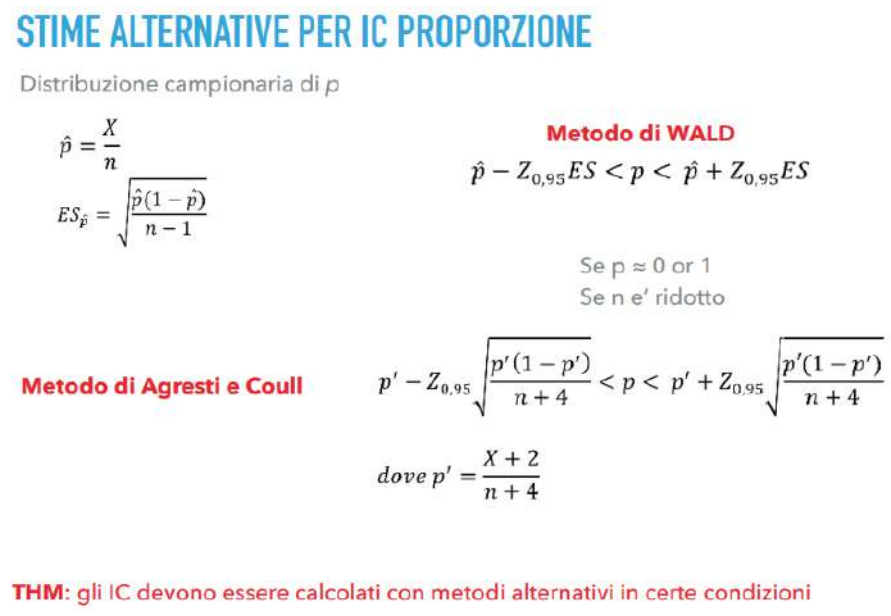
Il test binomiale fornisce un risultato ESATTO, ma il calcolo è molto lungo con il p value, soprattutto se n è grande. Esistono però delle alternative che permettono di svolgere l'analisi più rapidamente, arrivando a p values approssimati non esatti ma vicini al vero valore.



STIMA DELL'ERRORE STANDARD: di solito **non siamo in grado di calcolare la deviazione standard della proporzione della popolazione perché non conosciamo il parametro della popolazione p** (probabilità di successo). Siamo però in grado di **STIMARE L'ERRORE STANDARD, usando la stima della proporzione**

$$ES_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}$$

Se l'errore standard è basso, vuol dire che LA STIMA CAMPIONARIA p CAP è MOLTO VICINA ALLA PROPORZIONE p DELLA POPOLAZIONE.



METODO DI AGRESTI E COULL
L'intervallo di confidenza contiene i valori più plausibili del parametro che proviamo a stimare in base ai dati. Ci sono molti modi per calcolare un intervallo di confidenza approssimato per una proporzione, ma nessuno è esatto. Il metodo che si consiglia di usare è questo **metodo di Agresti e Coull**, che permette di calcolare p' e valutare l'intervallo di confidenza sulla base di p'. sapendo che per un intervallo di confidenza di 95% devo avere gli estremi di **Z pari a 1,96 e -1,96** allora posso trovare l'intervallo di confidenza.
Il metodo più usato è però il METODO DI WALD

INTERVALLI DI CONFIDENZA

SI CALCOLANO PER LE STIME

RIFERIMENTO A DISTRIBUZIONI DI 'ORIGINE' NORMALE

VERIFICA DIFFERENZE SIGNIFICATIVE (MANCANZA DI SOVRAPPOSIZIONE)

SOVRAPPOSIZIONE NON SIGNIFICA NON SIGNIFICATIVITA'

THM: gli IC vanno usati *cum grano salis*

ERRORI DI I E II TIPO

Errori dei test statistici

		REALTA'	
		H_0 VERA	H_0 FALSA
TEST	P≤0,05 RIFIUTO H_0	I TIPO	CORRETTA
	P>0,05 NON RIFIUTO H_0	CORRETTA	II TIPO (β)

se ERRORE I TIPO (STIMATO CON P VALUE) ≤ SOGLIA PREDEFINITA (α) –> si rifiuta H_0

1-β = POTENZA DEL TEST

THM: il test delle ipotesi e' un approccio piu' potente alla statistica inferenziale

Questo metodo NON si consiglia di usarlo perché è poco accurato, però delimita un intervallo attorno alla stima di p cap attraverso un multiplo con Z del suo errore standard, dove di nuovo **Z è 1,96 per un intervallo di confidenza del 95%.**

ESEMPIO DI COME SI USA L'INTERVALLO DI CONFIDENZA!!!!

ESERCIZIO IN CLASSE FIGLI MASCHI RADIOLOGI

I radiologi hanno meno figli maschi.
Studio su 87 radiologi: 30 figli maschi

La differenza e' realmente diversa da quanto ci si aspetterebbe?

DEFINIRE H_0
CALCOLARE SE TALE DIFFERENZA E' STATISTICAMENTE SIGNIFICATIVA USANDO IC

$$\hat{p} = \frac{X}{n}$$

$$p' - Z_{0,95}\sqrt{\frac{p'(1-p')}{n+4}} < p < p' + Z_{0,95}\sqrt{\frac{p'(1-p')}{n+4}}$$

$$\hat{p} - Z_{0,95}ES < p < \hat{p} + Z_{0,95}ES$$

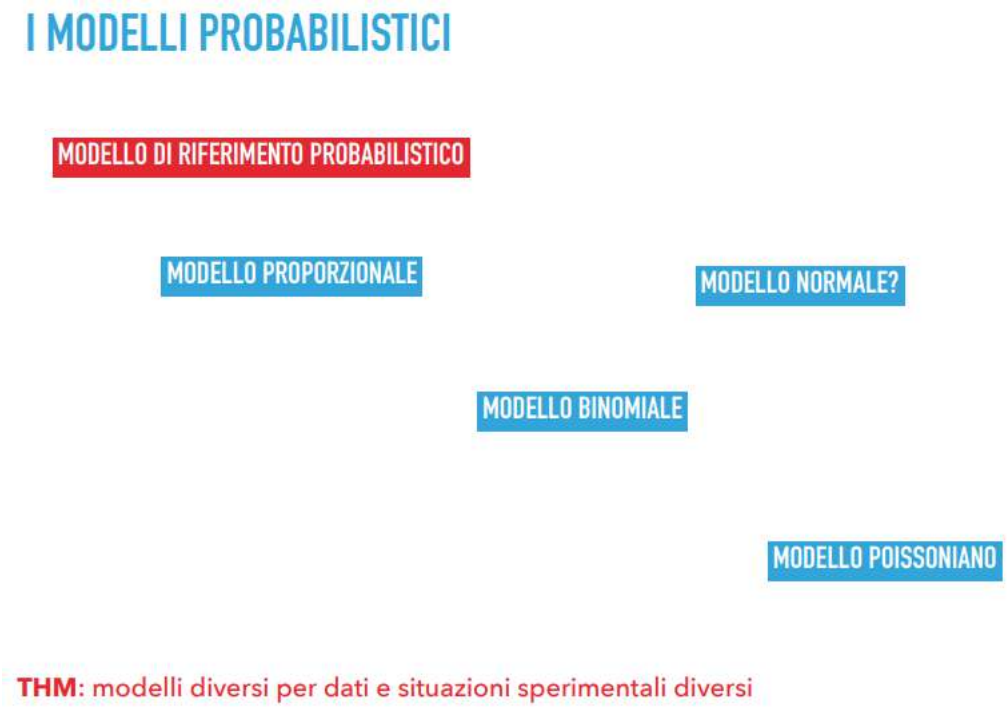
$$\Pr[X \text{ eventi}] = \binom{n}{X} p^{x}(1-p)^{n-x}$$

LAVORARE IN COPPIE
15 MINUTI

Esempio esercizio: sembra che i radiologi abbiano meno figli maschi. Lo studio prevede che su 87 radiologi ci siano 30 figli maschi. Sappiamo di avere una distribuzione binomiale, dove la distribuzione NULLA è data dal p=0,512, che è la proporzione di figli maschi nella popolazione umana
Ipotesi alternativa: i figli maschi non sono il 50%.
A questo punto

sapendo la distribuzione binomiale, so che X è 30. Così posso calcolare il p value. **Però la procedura è LUNGHISSIMA. Per questo è meglio usare l'INTERVALLO DI CONFIDENZA.**
Quindi ho X=30 e n=87, uso il metodo di Agresti e Coull usando Z=1,96. Svolgendo i calcoli trovo che il mio parametro deve stare tra 0,254 e 0,450, e questo **intervallo NON CONTIENE 0,512, CHE è LA PROPORZIONE DI FIGLI MASCHI CHE SI OSSERVA NELLA POPOLAZIONE UMANA! Quindi 0,512 NON è un valore plausibile. si può quindi RIFIUTARE L'IPOTESI NULLA**.

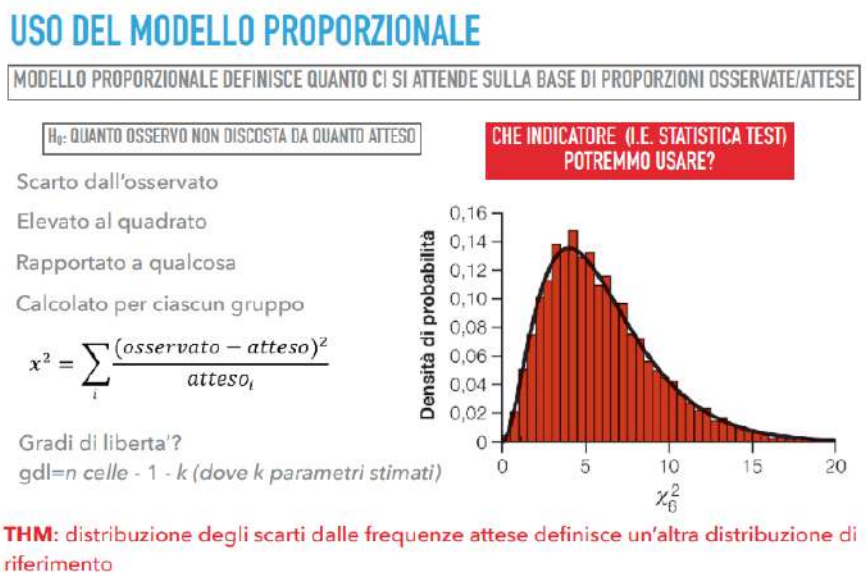
MODELLI PROBABILISTICI: TEST DEL CHI QUADRO



MODELLO PROPORZIONALE: TEST DEL CHI QUADRO

Il modello proporzionale è un modello probabilistico semplice in cui **LA FREQUENZA CON CUI SI VERIFICANO GLI EVENTI è PROPORZIONALE AL NUMERO DI VOLTE CHE HANNO LA POSSIBILITA' DI VERIFICARSI**.

Per esempio, si usa il modello proporzionale per dire che la frequenza dei geni della spermatogenesi sul cromosoma X del topo è direttamente proporzionale alle dimensioni del cromosoma. Per studiare il modello proporzionale viene usato il test del CHI QUADRO.



Il test del chi quadrato si usa per **misurare la discrepanza tra una distribuzione di frequenza osservata e le frequenze previste da un modello casuale che rappresenta l'ipotesi nulla**. È un **test di BONTA' DI ADATTAMENTO**, e fa un confronto.

Esempio: nascite nei giorni della settimana. Si passa da solo 33 nati la domenica a 63 nati il mercoledì e il giovedì. L'assunzione è il MODELLO PROPOZIONALE, che è l'ipotesi nulla, e ci aspettiamo che non ci sia differenza nei giorni della settimana.

IPOTESI NULLA: la probabilità di nascita è la stessa per ogni giorno della settimana

IPOTESI ALTERNATIVA: la probabilità di nascita non è la stessa per ogni giorno della settimana

Di nuovo, ipotesi nulla specifica e alternativa non specifica

L'ipotesi nulla è il MODELLO PROPORZIONALE, e si usa per generare la frequenza attesa di nati ogni giorno della settimana. Per esempio, nel 1999 ogni giorno della settimana è comparso 52 volte tranne il sabato che è comparso 53 volte. **QUINDI IL NUMERO DI NATI è PROPORZIONALE AL NUMERO DI VOLTE CHE IL GIORNO DELLA SETTIMANA C'E' STATO IN QUELL'ANNO**.

Così ottengo le proporzioni. Quindi se so che il numero totale di nati è 350, avrò per tutti i giorni della settimana tranne il venerdì (52/365)(350) nati.

LE FREQUENZE ATTESE POSSONO AVERE COMPONENTI ANCHE FRAZIONARIE, PERCHE' SONO I VALORI ATTESI.

Ovviamente la somma dei valori attesi deve essere uguale a quella dei valori osservati, quindi è bene arrotondare in modo che sia così.

A questo punto si fa la **STATISTICA TEST CHI QUADRO**: "osservato" è la frequenza di individui osservata nella i-esima categoria, "atteso" quella del modello proporzionale nella stessa categoria (sotto l'ipotesi nulla). **LA STATISTICA CHI QUADRO USA LE FREQUENZE ASSOLUTE, NON QUELLE RELATIVE (AKA NON USA LE PROPORZIONI)**

Si ottiene il chi quadro finale. **Si ottiene come numero 15,05: è sufficiente per rifiutare l'ipotesi nulla**?

In basso a sinistra c'è il grafico che mostra la distribuzione dei valori del chi quadro fatta al computer tenendo conto del modello proporzionale dell'ipotesi nulla. È però possibile usare un metodo più semplice che non sia basato sulle analisi al computer. La distribuzione nulla è infatti ben approssimata dalla distribuzione TEORICA del chi quadro, che è matematicamente nota (ed è sempre quella in basso a sinistra o destra). Essendo sempre quelle, **le caratteristiche della distribuzione del chi quadro sono state TABULATE. Quindi la distribuzione chi quadro è una FAMIGLIA DI DISTRIBUZIONI, e si può capire quale distribuzione della tabella usare in base al numero di gradi di libertà**.

Il numero di gradi di libertà specifica quale distribuzione usare e si calcola facendo (numero di categorie) – 1 – (numero di parametri stimato in base ai dati). In questo caso l'ultimo termine è 0 (poi vedremo perché), quindi il numero di gradi di libertà totali è 6.

QUINDI OGNI VOLTA CHE HO UNA IPOTESI NULLA CON MODELLO PROPORZIONALE POSSO USARE IL CHI QUADRO!!!! CHE è GIA' TABULATO.

Quindi cerchiamo la distribuzione (chi quadro)(6): è molto simile alla distribuzione (chi quadro)(1), ma la prima è un pochino più regolare.

USIAMO QUESTA DISTRIBUZIONE PER CALCOLARE IL P VALUE DEL TEST

Se i dati fossero esattamente corrispondenti all'ipotesi nulla, il chi quadro sarebbe 0: si usa quindi un TEST A UNA CODA: **quindi la probabilità di ottenere un valore di chi quadro maggiore di 15,05 è data dall'area sottesa alla curva del chi quadro oltre 15,05**.

Come troviamo l'area? Si usano tabelle già fatte, e si trova che P=0,0199, che ci fa RIFIUTARE L'IPOTESI NULLA!

Per determinare il p value si possono anche usare i VALORI CRITICI

Se per esempio usiamo un livello di significatività di 0,05, SAPPIAMO PER OGNI NUMERO DI GRADI DI LIBERTA' A COSA CORRISPONDE UN'AREA DI 0,05!!!! Quindi se il nostro chi quadro per quei gradi di libertà è inferiore al valore che dà 0,05 vuol dire che NON possiamo rifiutare l'ipotesi nulla, viceversa la rifiutiamo.

Vedi nella tabella per 15,05: prendo 6 gradi di libertà, e vedo che per avere 0,05 di area ho 12,59. Visto che 15,05 è più alto, vuol dire che rifiuto l'ipotesi nulla.

	Livello di significatività α									
df	0,999	0,995	0,99	0,975	0,95	0,05	0,025	0,01	0,005	0,001
1	0,0000002	0,00004	0,00016	0,00098	0,00393	3,84	5,02	6,63	7,88	10,83
2	0,002	0,01	0,02	0,05	0,10	5,99	7,38	9,21	10,6	13,82
3	0,02	0,07	0,11	0,22	0,35	7,81	9,35	11,34	12,84	16,27
4	0,09	0,21	0,30	0,48	0,71	9,49	11,14	13,28	14,86	18,47
5	0,21	0,41	0,55	0,83	1,15	11,07	12,83	15,09	16,75	20,52
6	0,38	0,68	0,87	1,24	1,64	12,59	14,45	16,81	18,55	22,46
7	0,60	0,99	1,24	1,69	2,17	14,07	16,01	18,48	20,28	24,32
8	0,86	1,34	1,65	2,18	2,73	15,51	17,53	20,09	21,95	26,12

ESERCIZIO: VERIFICA DEL MODELLO PROPORZIONALE

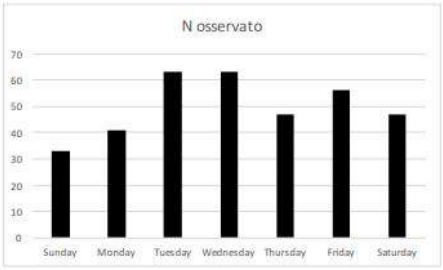
NASCITE E GIORNI DELLA SETTIMANA

Day	N
Sunday	33
Monday	41
Tuesday	63
Wednesday	63
Thursday	47
Friday	56
Saturday	47

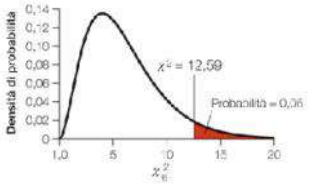
NON
LE NASCITE SONO DISTRIBUITE UNIFORMEMENTE NEI VARI GIORNI DELLA SETTIMANA

52 GIORNI DELLA SETTIMANA E 53 SABATI

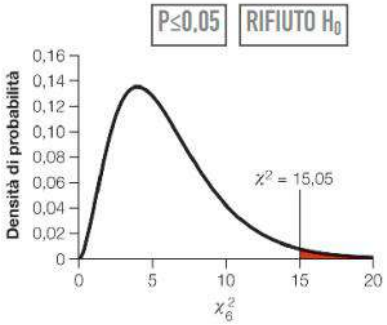
?



$$x^2 = \sum_i \frac{(osservato - atteso)^2}{atteso_i}$$



P≤0,05 RIFIUTO H_0



ASSUNZIONI DEL CHI QUADRO

Questo test assume che gli individui nell'insieme dei dati siano un campione casuale estratto dall'intera popolazione. Questo vuol dire che ogni individuo ha la stessa probabilità degli altri di essere incluso nella popolazione.

Il chi quadro fornisce comunque dei valori approssimati. Perché l'approssimazione sia fatta bene, è necessario che **nessuna delle categorie scelte abbia una frequenza attesa minore di 1, e che non più del 20% delle categorie abbia una frequenza attesa minore di 5**. Ovviamente questo si riferisce alle frequenze ATTESE, NON osservate. Se queste condizioni non sono soddisfatte, il test diventa inaffidabile.

Test del chi quadro con SOLO DUE CATEGORIE

Può essere usato in SOSTITUZIONE AL TEST BINOMIALE, perché i calcoli sono molto più veloci, anche se sono meno esatti.

Esempio: ereditarietà dei cromosomi sessuali. Ci aspettiamo che il 5,2% dei geni sia localizzato sul cromosoma X, visto che contiene quella percentuale di materiale genetico totale

IPOTESI NULLA: i geni umani sul cromosoma X sono il 5,2% del totale

IPOTESI ALTERNATIVA: i geni umani sul cromosoma X NON sono il 5,2% del totale: TEST A DUE CODE.

Il numero atteso di geni è 20290x0,052=1055, ma ne sono stati osservati 781.

Come faccio a capire se posso non rifiutare l'ipotesi nulla?

Calcolare il p value usando il test binomiale è motlo complicato, perché abbiamo troppi numeri! Dovremmo infatti calcolare 2P(X<=781)

Si può però fare più rapidamente usando il test del chi quadro. Ho **DUE CATEGORIE: GENI SUL CROMOSOMA X E GENI NON SUL CROMOSOMA X**

$$\chi^2 = \frac{(781 - 1055)^2}{1055} + \frac{(19\,509 - 19\,235)^2}{19\,235} = 75{,}1$$

Questa statistica test ha due categorie e, quindi, soltanto un grado di libertà:

$$df = 2 - 1 = 1$$

Vediamo dalla tavola che ho messo sopra che il valore critico di chi quadro per una significatività paria a 0,05 è 3,84. Quindi posso RIFIUTARE L'IPOTESI NULLA.

QUINDI CON DUE CATEGORIE IL TEST BINOMIALE è PIU' ADATTO SE N è PICCOLO, PERCHE' È PRECISO, MENTRE IL CHI QUADRO QUANDO N è GRANDE.

ESERCIZIO: VERIFICA DEL MODELLO PROPORZIONALE

CROMOSOMA X E CONTENUTO GENICO

CONTIENE 5,2% DEL DNA TOTALE
CONTIENE 781 GENI SU 20290

H_0: IL CROMOSOMA X NON CONTIENE MENO GENI DI ALTRI

?

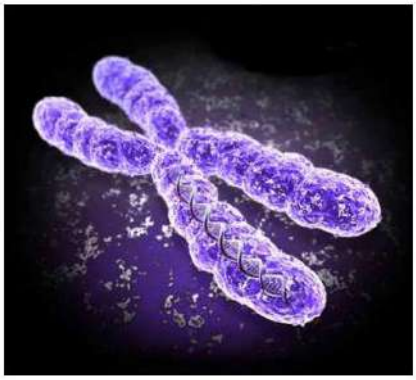
$$x^2 = \sum_i \frac{(osservato - atteso)^2}{atteso_i}$$

X^2=75,1; GDL=2-1=1

$X^2_{(1;0,05)}$=3,84; GDL=1

$X^2_{(1;0,001)}$=10,83; GDL=1

P≤? P≤0,001 RIFIUTO H_0



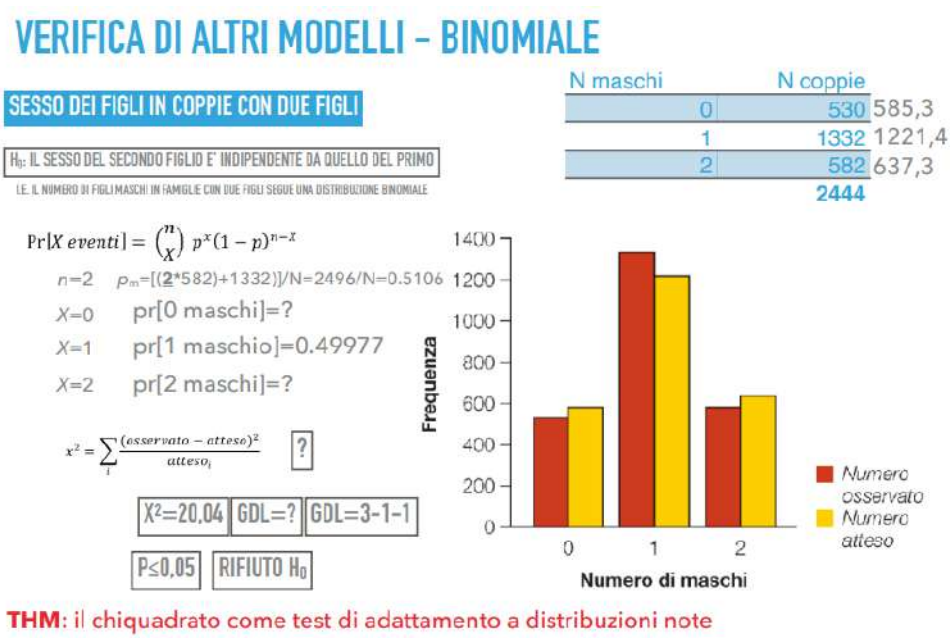
Di solito quando si vuole descrivere qualcosa che è successa in natura si cerca di capire quale modello è il migliore. **Tra i modelli statistici più usati abbiamo quello PROPORZIONALE, BINOMIALE E POISSONIANO: MODELLI DIVERSI PER DATI E SITUAZIONI SPERIMENTALI DIVERSE**.
La distribuzione BINOMIALE e quella di POISSON descrivono entrambe la probabilità di ottenere un certo numero di successi in prove indipendenti. **La distribuzione binomiale indica la possibilità di ottenere X successi in n prove indipendenti, quella di Poisson di avere X successi in un certo intervallo di tempo o in una certa porzione di spazio**.
Esempio: il fatto di aver generato un figlio maschio non influenza il sesso del secondogenito. Ipotizziamo quindi che il numero di maschi e di femmine in famiglie con due figli sia una BINOMIALE, con n=2 e p=probabilità di avere un maschio in n prove, e **vogliamo capire se è la rappresentazione migliore**.
IPOTESI NULLA: il numero di maschi in famiglie con due figli ha una distribuzione binomiale
IPOTESI ALTERNATIVA: non ha una distribuzione binomiale
IN UN TEST BINOMIALE ABBIAMO UN SOLO SET DI PROVE, non stiamo cercando, come prima nella distribuzione binomiale, di verificare un'ipotesi nulla su una singola proporzione. Inoltre, l'ipotesi nulla non specifica p, la probabilità di avere un singolo figlio maschio. Quindi dobbiamo STIMARE p in base ai dati. Sappiamo che ci sono 4888 figli, valore ottenuto perché abbiamo 2444 famiglie, ciascuna con 2 figli. il numero totale di maschi OSSERVATO è stabilito in base ai dati che abbiamo, ed è 2x582 (aka numero di famiglie con due maschi) + 1332 (aka numero di famiglie con un maschio)
Usiamo ora questa p per fare la binomiale
LA COSA CHE STIAMO FACENDO ORA, INFATTI, NON è VERIFICARE CHE I MASCHI E LE FEMMINE SI DISTRIBUISCANO UGUALI NELLA POPOLAZIONE, MA CHE IL TEST BINOMIALE SIA IL MIGLIORE!!!! ASSUMENDO CHE DATI SIANO SENSATI!!! E VEDENDO SE IL BINOMIALE è MEGLIO DEL POISSON O VICEVERSA. PER QUESTO USIAMO I DATI REALI PER CREARE LA BINOMIALE, PERCHE' VOGLIAMO CAPIRE SE LA BINOMIALE PUO' POI ESSERE USATA PER CAPIRE SE DAVVERO IL NUMERO DI F E M è UGUALE, MA QUESTO DOPO!!!
Quindi uso questa p per trovare la distribuzione binomiale, e cerco la P di avere un maschio come

$$\Pr[1\ maschio] = \binom{2}{1}(0{,}5106)^1(1-0{,}5106)^1 = 0{,}49977$$

E quindi CON LA BINOMIALE la **frequenza ATTESA** di 2444 famiglie con 2 figli che hanno un maschio

$Atteso\ (1\ maschio) = 2444 \times 0{,}49977 = 1221{,}4$

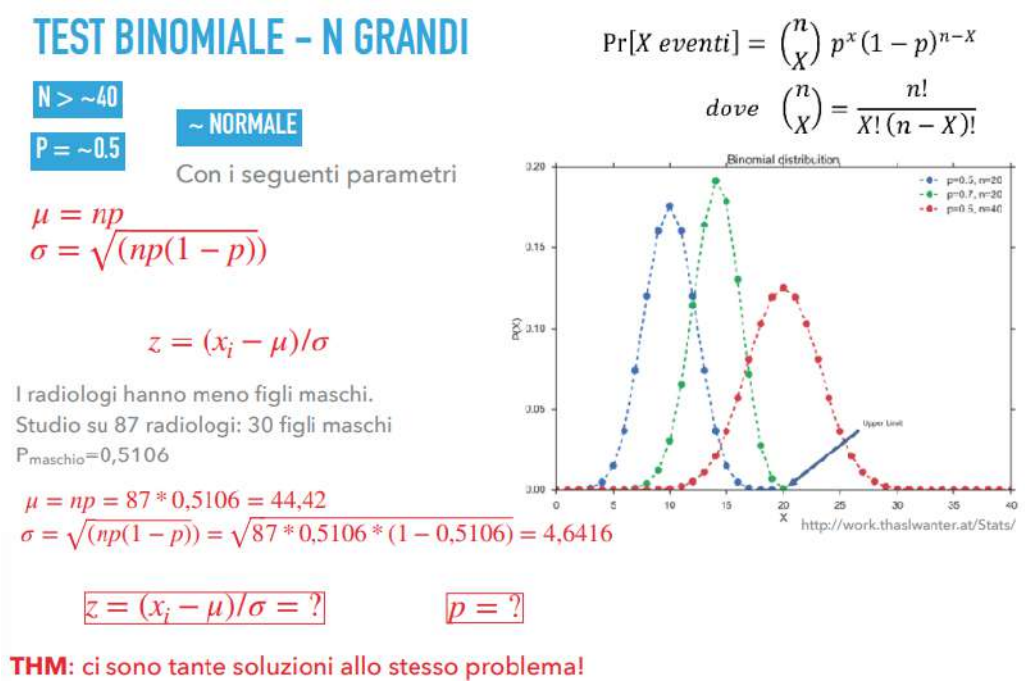
Posso fare la stessa cosa con due maschi e zero maschi. Quindi si vede **che LE FREQUENZE OSSERVATE NON SONO UGUALI A QUELLE ATTESE,** anche se le differenze non sono enormi
Usiamo il chi quadro. Otteniamo che il chi quadro è 20,04. Sapendo che i gradi di libertà sono **3 categorie – 1 – PARAMETRO DI DATI STIMATO PER CALCOLARE LE FREQUENZE ATTESE**: quindi ho 1 grado di libertà!



Il valore per p=0,05 è 3,84, quindi di nuovo posso rifiutare l'ipotesi nulla. Se però esploro di più la tavola statistica, vedo che 20,04 è maggiore anche del valore critico corrispondente a 0,001. QUESTO CI DICE CHE LA DISTRIBUZIONE DI FREQUENZA DEL NUMERO DI MASCHI E FEMMINE DELLE FAMIGLIE CON 2 FIGLI NON è GOVERNATO DA UNA DISTRIBUZIONE BINOMIALE. DEVO RIFIUTARE LA BINOMIALE.

Come mai la binomiale non si adatta? Forse perché le famiglie con due figli senza maschi o con due maschi decidano di fare il terzo figlio per avere un figlio del sesso opposto e così non rientrano nella statistica. Se no non si spiega, perché la distribuzione dovrebbe essere binomiale.

Da qui possiamo pensare che la distribuzione migliore sia di POISSON



Altra cosa importante prima di passare alla distribuzione di Poisson: una BINOMIALE CON N GRANDE SI APPROSSIMA AD UNA GAUSSIANA!

Ora passiamo al **modello di Poisson**

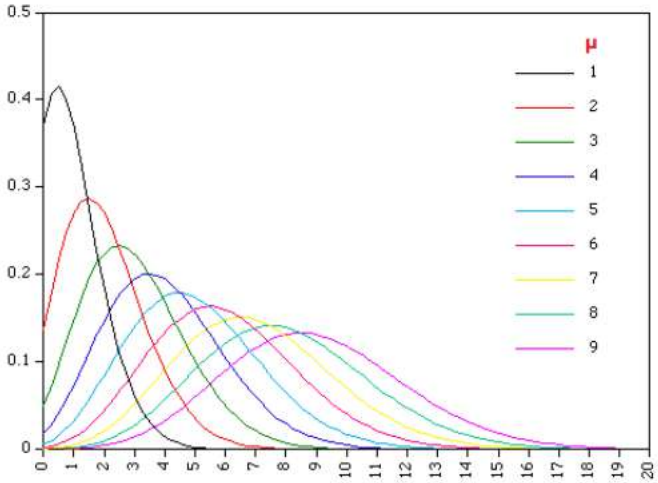
MODELLO POISSONIANO

NUMERO DI SUCCESSI NELLO SPAZIO E/O NEL TEMPO

$$\Pr[x\ successi] = \frac{e^{-\mu}\mu^{X}}{X!}$$

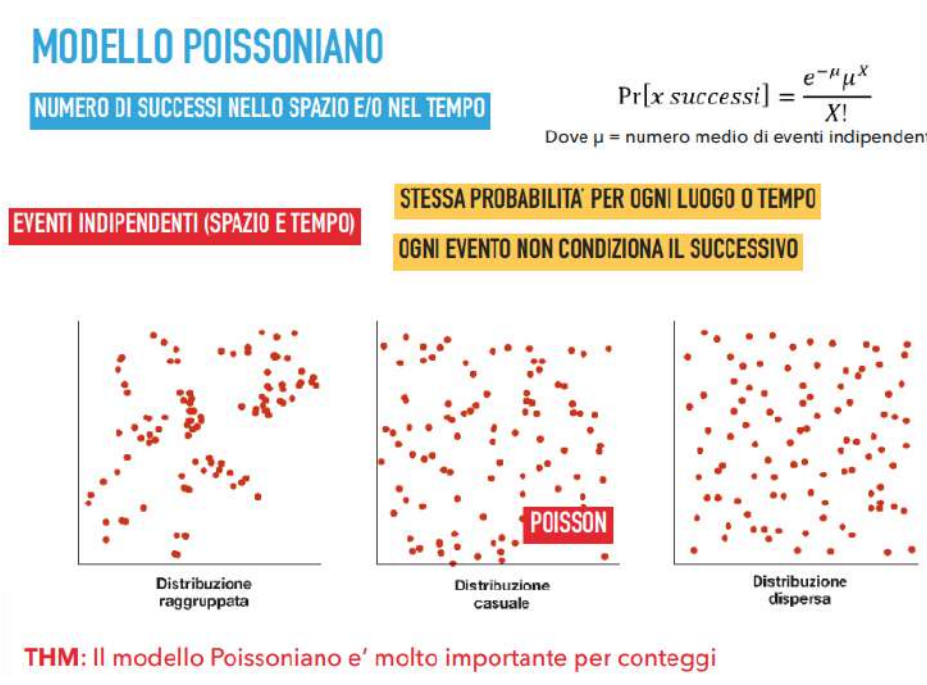
Dove μ = numero medio di eventi indipendenti in S e T

EVENTI INDIPENDENTI (SPAZIO E TEMPO)



Esempio: polvere dovuta ad un'eruzione che si deposita sul terreno. I primi organismi a colonizzare il territorio circostante di nuovo furono i ragni, portati dalle correnti di aria, e aumentano in numero alimentandosi di insetti arrivati nello stesso modo. Come arrivano i ragni sul suolo? Come sarebbe la loro distribuzione se arrivassero in modo casuale?

Assunzioni: la probabilità che un ragno atterri in un dato punto del terreno è la stessa ovunque, E il fatto che un ragno atterra in un dato punto è indipendente da dove sono atterrati gli altri. Se entrambi gli assunti sono soddisfatti, la distribuzione che prendiamo in considerazione è una POISSONIANA.



Quindi una distribuzione di Poisson descrive il **NUMERO DI SUCCESSI IN INTERVALLI SPAZIALI O TEMPORALI QUANDO SI VERIFICANO INDIPENDENTEMENTE L'UNO DALL'ALTRO E CON UGUALE PROBABILITA' IN OGNI PUNTO DEL TEMPO O DELLO SPAZIO**. CI PERMETTE DI CAPIRE SE GLI EVENTI AVVENGONO CASUALMENTE NEL TEMPO O NELLO SPAZIO.

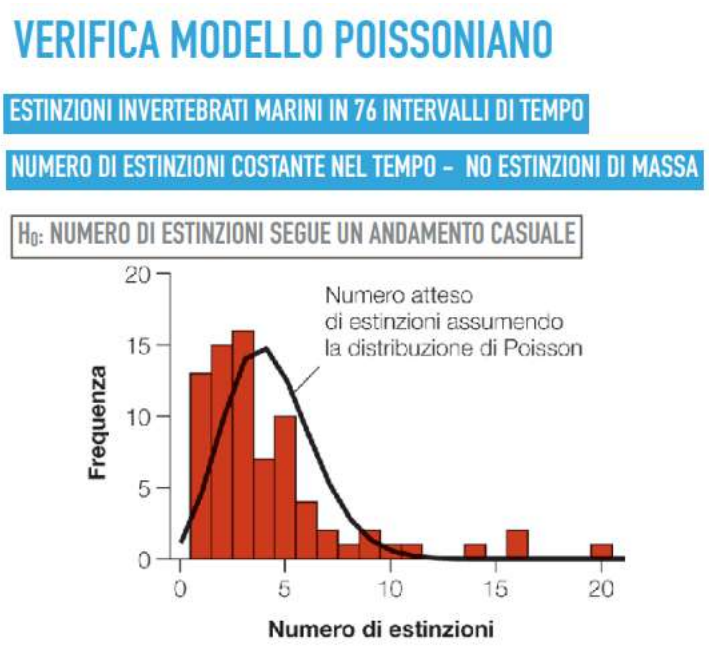
Ci sono molte cose in biologia per cui ci si aspetta una poissoniana: il numero di salmoni catturati in un certo giorno da pescatori sportivi o il numero di semi che germinano per ogni pianta. ovviamente è sempre un MODELLO.

Ovviamente è possibile che i **successi siano distribuiti nello spazio o nel tempo in modo non casuale, quindi RAGGRUPPATO o DISPERSO**. Per esempio, le epidemie di malattie contagiose hanno spesso una

distribuzione spaziale raggruppata, perché gli individui malati infettano individui vicini. Al contrario, una distribuzione dispersa si ha quando un successo rende più difficile ad un secondo successo di accadere.
Si vede nella slide la formula della distribuzione di Poisson

VERIFICA DEL MODELLO POISSONIANO

In biologia il modello poissoniano è usato nella formulazione di una ipotesi nulla che **verifichi che i successi siano distribuiti CASUALMENTE** nello spazio o nel tempo.



N	Frequency
0	0
1	13
2	15
3	16
4	7
5	10
6	4
7	2
8	1
9	2
10	1
11	1
12	0
13	0
14	1
15	0
16	2
17	0
18	0
19	0
20	1

Esempio: estinzioni degli invertebrati marini in 76 intervalli di tempo di durata simile. **Se l'estinzione fosse casuale nel tempo, allora il numero di estinzioni in ogni intervallo di tempo dovrebbe essere una poissoniana**.
Oppure potrebbero essere distribuite in modo più raggruppato per qualche motivo.
Il metodo è sempre usare il CHI QUADRO.
IPOTESI NULLA: il numero di estinzioni negli intervalli di tempo ha una distribuzione di Poisson
IPOTESI ALTERNATIVA: non ha una distribuzione di Poisson
Inizialmente dobbiamo stimare il numero medio di estinzioni per ogni intervallo di tempo, con **76 gli intervalli di tempo e N il numero di estinzioni ogni intervallo** (il totale della seconda colonna è 76).

$$\bar{X} = \frac{(0 \times 0) + (13 \times 1) + (15 \times 2) + \ldots}{76} = 4{,}21$$

L'istogramma è la distribuzione di frequenza osservata, mentre la funzione esprime il numero atteso seguendo la distribuzione di Poisson.

Per capire se si può usare Poisson, usiamo il test del Chi quadro.
Per calcolare le frequenze attese basta usare Poisson per ogni categoria di estinzione. Moltiplichiamo poi il valore di Poisson per il numero totale di intervalli di tempo

$$\Pr[3\ \textit{estinzioni}] = \frac{e^{-\mu}\mu^3}{3!} = \frac{e^{-4{,}21}(4{,}21)^3}{3!} = 0{,}1846$$

Moltiplicando questo risultato per il numero totale di intervalli di tempo nell'insieme dei dati (76), otteniamo il numero atteso di intervalli di tempo con 3 estinzioni:

$$\textit{Atteso}[3\ \textit{estinzioni}] = 76 \times 0{,}1846 = 14{,}03$$

VERIFICA MODELLO POISSONIANO

ESTINZIONI INVERTEBRATI MARINI IN 76 INTERVALLI DI TEMPO

NUMERO DI ESTINZIONI COSTANTE NEL TEMPO – NO ESTINZIONI DI MASSA

H_0: NUMERO DI ESTINZIONI SEGUE UN ANDAMENTO CASUALE

COME FARE?

$$\Pr[x\ successi] = \frac{e^{-\mu}\mu^{X}}{X!}$$

Dove μ = numero medio di eventi indipendenti in S e T

$$\Pr[3\ successi] = \frac{e^{-\mu}\mu^{3}}{3!} = \frac{e^{-4{,}21}(4{,}21)^{3}}{3!} = 0{,}1846$$

$$\text{Atteso}[3\ successi] = 76 \times 0{,}1846 = 14{,}03$$

N	Frequency	Expected
0	0	1.13
1	13	4.75
2	15	10.00
3	16	14.03
4	7	14.77
5	10	12.44
6	4	8.72
7	2	5.24
8	1	2.76
9	2	1.29
10	1	<1
11	1	<1
12	0	<1
13	0	<1
14	1	<1
15	0	<1
16	2	<1
17	0	<1
18	0	<1
19	0	<1
20	1	<1

Ora, calcolando il chi quadro si vede che i valori non corrispondono. DEVO QUINDI RIFIUTARE L'IPOTESI NULLA.

Ps: per i gradi di libertà di nuovo devo togliere 1 perché ho stimato

VERIFICA MODELLO POISSONIANO

ESTINZIONI INVERTEBRATI MARINI IN 76 INTERVALLI DI TEMPO

NUMERO DI ESTINZIONI COSTANTE NEL TEMPO – NO ESTINZIONI DI MASSA

~~H_0: NUMERO DI ESTINZIONI SEGUE UN ANDAMENTO CASUALE~~ NON

N	Observed	Expected
0	0	1.13
1	13	4.75
2	15	10.00
3	16	14.03
4	7	14.77
5	10	12.44
6	4	8.72
7	2	5.24
8	1	2.76
9	2	1.29
10	1	<1
11	1	<1
12	0	<1
13	0	<1
14	1	<1
15	0	<1
16	2	<1
17	0	<1
18	0	<1
19	0	<1
20	1	<1

N	Observed	Expected
0-1	13	5.88
2	15	10.00
3	16	14.03
4	7	14.77
5	10	12.44
6	4	8.72
7	2	5.24
>=8	9	4.91

X^2=23,93 | GDL=? | GDL=8-1-1=6

P≤0.001 | RIFIUTO H_0

VERIFICA MODELLO POISSONIANO ET AL – VMR

RAPPORTO VARIANZA/MEDIA | VARIANCE-TO-MEAN RATIO (VMR) $VMR = D = \sigma^2/\mu$

VMR > 1 DISTRIBUZIONE DISPERSA

VMR < 1 DISTRIBUZIONE RAGGRUPPATA

VMR ~= 1 ? ← POISSON

https://en.wikipedia.org/wiki/Index_of_dispersion

Distribution	VMR	
constant random variable	VMR = 0	not dispersed
binomial distribution	0 < VMR < 1	under-dispersed
Poisson distribution	VMR = 1	RANDOM
negative binomial distribution	VMR > 1	over-dispersed

THM: rapporto varianza/media molto semplice e utile per comprendere distribuzione serie di dati

ASSOCIAZIONE TRA VARIABILI CATEGORICHE BINARIE

ASSOCIAZIONE TRA VARIABILI CATEGORICHE BINARIE

		Variabile X1	(colonne; c)
		C	D
Variabile X2	A	*nAC*	*nAD*
(righe; r)	B	*nBC*	*nBD*

INDIPENDENTI?

SE INDIPENDENZA — ATTESO — PR[A E C]=PR[A] X PR[C]

SE NON INDIPENDENZA — -> — PR[A E C]≠PR[A] X PR[C]

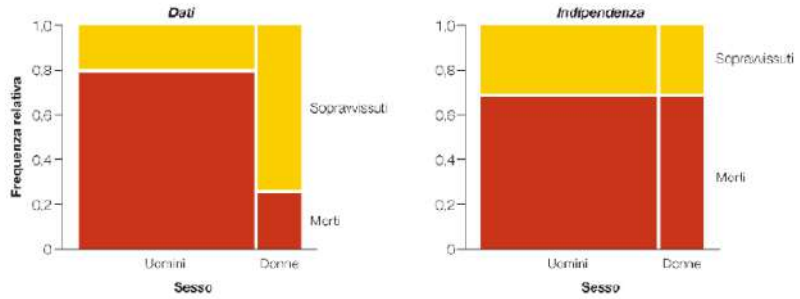
THM: la teoria della probabilita' aiuta a comprendere l'atteso

Due variabili categoriche associate NON SONO INDIPENDENTI. Ovviamente se non sono indipendenti la probabilità dell'intersezione NON è il prodotto delle probabilità

ASSOCIAZIONE TRA VARIABILI CATEGORICHE BINARIE

MORTALITA' TITANIC PER SESSO

	deceduti	sopravvissuti	
M	1364	367	*1731*
F	126	344	*470*
	1490	*711*	**2201**



THM: associazione tra variabile categorica 'dipendente' e variabile categorica 'indipendente'

Esempio: titanic. Le donne avevano una probabilità di morire inferiore degli uomini. In quella notte, quindi, SESSO E MORTALITA' NON ERANO INDIPENDENTI. Questo si può vedere nel **MOSAIC PLOT**: se la mortalità fosse stata indipendente dal sesso la probabilità di morire sarebbe stata la stessa, e si avrebbe ottenuto il MOSAIC PLOT a destra.

ASSOCIAZIONE TRA VARIABILI CATEGORICHE BINARIE: RR

		Variabile X1	(colonne; c)
		C	D
Variabile X2	A	*nAC*	*nAD*
(righe; r)	B	*nBC*	*nBD*

RISCHIO RELATIVO $\widehat{RR} = \frac{p_{event|A}}{p_{event|B}}$ Dove l'*evento* è il fenomeno osservato e A e B sono i due gruppi di confronto

Capitolo 9.2

RIDUZIONE RISCHIO RELATIVO $1 - \widehat{RR} = 1 - \frac{p_{event|A}}{p_{event|B}}$

RIDUZIONE RISCHIO ASSOLUTO $= p_{event|A} - p_{event|B}$

THM: il rischio relativo è il rapporto tra le probabilità che un evento accada in due gruppi di confronto, serve per valutare quanto cambi il rischio in due condizioni differenti.

STIMA DELL'ASSOCIAZIONE

RISCHIO RELATIVO: **quanto è rischioso l'evento nel gruppo A rispetto a quanto è rischioso lo stesso evento nel gruppo B**? aka probabilità di morte in A fratto in B. SERVE PER VALUTARE QUANTO CAMBIA IL RISCHIO IN DUE CONDIZIONI DIVERSE.

Quando abbiamo DUE variabili associate vogliamo capire QUANTO SONO ASSOCIATE. Per questo, usiamo l'ODDS RATIO, ma solo quando ciascuna delle due variabili ha solo due categorie. In questo caso, la variabile di **RISPOSTA** è quella che ha **come categorie SUCCESSO e INSUCCESSO**. L'altra variabile è invece quella **ESPLICATIVA**, in cui le categorie sono i **due gruppi in cui vengono CALCOLATE E CONFRONTATE LE PROBABILITA' DI SUCCESSO**.

Per convenzione, chiamiamo con **p la probabilità di successo, e con 1-p la probabilità di insuccesso**. L'odds di un successo è p/(1-p).

Se O=1 si verifica un successo per ogni insuccesso. Se O=10 invece si verificano 10 successi ogni insuccesso.

Si parla di **ODDS DI SUCCESSO**: probabilità di successo divisa per la probabilità di insuccesso.

ASSOCIAZIONE TRA VARIABILI CATEGORICHE BINARIE

		Variabile X1	(colonne; c)
		C	D
Variabile X2	A	*nAC*	*nAD*
(righe; r)	B	*nBC*	*nBD*

ODDS RATIO Rapporto tra gli 'odds' (*inequalities; degree of unlikeness*)

$$ODDS = \frac{p}{(1-p)} \quad \text{e.g.,} \quad ODDS_A = \frac{p_A}{(1-p_A)}$$

$$ODDS_{ratio} = \frac{p_{A|C}}{(1-p_{A|C})} / \frac{p_{A|D}}{(1-p_{A|D})}$$

Se = 1 -> nessuna associazione

Se ≠ 1 -> associazione

THM: gli odds ratio confrontano il cambiamento delle probabilità di ottenere una certa risposta dicotomica al variare di un'altra variabile dicotomica

Allo stesso modo, si può fare la stima dell'odds di successo usando la proporzione di successo osservata

$$\hat{O} = \frac{\hat{p}}{1-\hat{p}}$$

Questa formula si usa se si usano DATI SPERIMENTALI e p cap è un rapporto.

Possiamo poi fare O1 e O2, dove O1 identifica il primo gruppo, O2 il secondo gruppo

L'ODDS RATIO è il RAPPORTO TRA L'ODDS DI SUCCESSO DI UN GRUPPO E L'ODDS DI SUCCESSO DEL SECONDO GRUPPO. Se l'odds ratio è 1, vuol dire che l'odds di successo nella variabile di risposta è lo stesso per entrambi i gruppi. Se invece è maggiore di 1, vuol dire che l'evento ha un odds maggiore nel primo gruppo che nel secondo. Se invece è minore di 1 vuol dire che l'odds è maggiore nel secondo gruppo.

SE L'ODDS RATIO è 1 VUOL DIRE CHE NON C'E' ASSOCIAZIONE: non c'è differenza nel successo/insuccesso dei due gruppi.

SE C'E' L'ACCENTO CIRCONFLESSO SULLA O VUOL DIRE CHE è UNA STIMA DELL'OR NELLA POPOLAZIONE DI INTERESSE.

ASSOCIAZIONE TRA VARIABILI CATEGORICHE BINARIE

DONNE		Trattament	(10 anni)
		Aspirina	Placebo
Cancro	SI	a=1438	b=1427
	NO	c=18496	d=18515

SONO INDIPENDENTI?

I.E., L'ASPIRINA RIDUCE IL RISCHIO DI SVILUPPARE UN CANCRO INVASIVO?

$$ODDS_{ratio} = \frac{p_{A|C}}{(1-p_{A|C})} / \frac{p_{A|D}}{(1-p_{A|D})}$$

STIMA

Se = 1 -> nessuna associazione

Se ≠ 1 -> associazione

PROBLEMA?

INTERVALLO DI CONFIDENZA DELL'ODDS RATIO

$$ES[\ln(\widehat{OR})] = \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$$

$$\ln(\widehat{OR}) - zES[\ln(\widehat{OR})] < \ln(OR) < \ln(\widehat{OR}) + zES[\ln(\widehat{OR})]$$

$$e^{LimINF} < OR < e^{LimSUP}$$

CALCOLARE GLI ODDS/RATIO!

THM: gli IC degli ODDS RATIO si usano per verificare la significativita' dell'associazione tra variabili dicotomiche

A Mascolo, Dip Biologia, Univ Pisa | alessandro.mascolo@unipi

La distribuzione campionaria dell'odds ratio è molto asimmetrica, motivo per cui è meglio convertire l'odds ratio nel suo logaritmo naturale, e solo a questo punto calcolarne l'errore standard (vedi slide)
Da qui si può anche calcolare l'intervallo di confidenza approssimato per il logaritmo naturale dell'odds ratio. Una volta ottenuta la diseguaglianza basta elevare su base e gli estremi dell'intervallo. Se gli estremi dell'intervallo di confidenza sono centrati su 1 vuol dire che è molto difficile che ci sia correlazione.

RR O OR?

RISCHIO RELATIVO (RR) — PIÙ SEMPLICI DA INTERPRETARE

ODDS RATIO (OR) — QUANDO NON SI PUÒ STIMARE IL RISCHIO

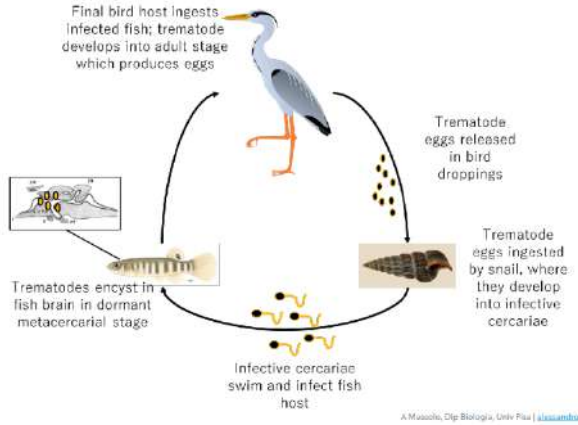
ESPERIMENTI CASO-CONTROLLO

CAMPIONAMENTO NON ADEGUATO

RISCHIO RELATIVO: so che c'è associazione e voglio sapere di quanto
ODDS RATIO: voglio sapere se l'associazione è significativa

THM: il rischio relativo si usa quanto si possono stimare i rischi con accettabile precisione, mentre gli ODDS RATIO si usano in caso contrario

CICLO EUHAPLORCHIS CALIFORNIENSIS



A Mascolo, Dip Biologia, Univ Pisa | alessandro.mascolo@unipi

Esempio: analisi delle tabelle di contingenza con il test del chi quadro.
IL TEST PIU' USATO PER L'ANALISI DELLE TABELLE DI CONTINGENZA PER STUDIARE L'ASSOCIAZIONE TRA DUE VARIABILI CATEGORICHE è IL TEST DEL CHI QUADRO!!!!!!!!!!!!!!!!!
ANCHE SE LE VARIABILI NON SONO BINARIEE (ma anche binarie va bene)

CHI-QUADRATO

PARASSITA EUHAPLORCHIS CALIFORNIENSIS
MANIPOLAZIONE DEL COMPORTAMENTO OSPITE (PESCE)

ALTERED BEHAVIOR OF PARASITIZED KILLIFISH INCREASES SUSCEPTIBILITY TO PREDATION BY BIRD FINAL HOSTS[1]

KEVIN D. LAFFERTY AND A. KIMO MORRIS[1]

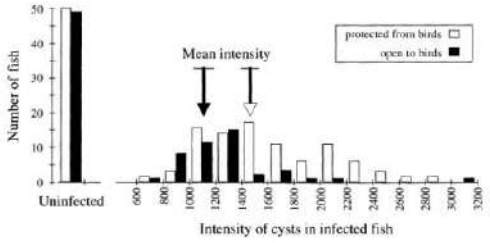


FIG. 2. The effect of bird predation on parasitized fish. The histogram shows the estimated frequency of infection intensities of the control (protected) and treatment (open) pens at the end of 20 d, indicating that parasitized fish were more likely to be eaten than unparasitized fish and that highly parasitized fish were more likely to be eaten than lightly parasitized fish.

Parassita: il trematode con quel nome impronunciabile usa tre ospiti durante il suo ciclo biologico: raggiunge la maturità negli uccelli, raggiungono l'ambiente esterno attraverso le feci dell'ospite. le uova sono poi ingerite da un gasteropode, che non riesce più a riprodursi. Quando viene poi ingerito da un pesce, il parassita si sviluppa e si incista nella scatola cranica del pesce fino a che non viene ingerito da un uccello. Diventa quindi un parassita adulto e riprende il ciclo. si pensa che l'infestazione influenza il rischio di predazione da parte degli uccelli.

Si prendono uccelli non infestati, lievemente infestati e infestati. Si riporta il numero di pesci ingeriti a seconda di quanto erano infestati, perché potrebbe esserci correlazione tra il fatto che i pesci sono infestati con il fatto che vengano predati di più

CHI-QUADRATO

PARASSITA EUHAPLORCHIS CALIFORNIENSIS
MANIPOLAZIONE DEL COMPORTAMENTO OSPITE (PESCE)

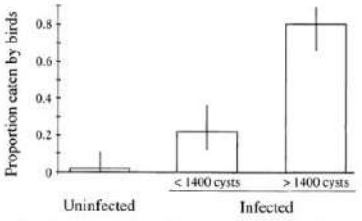
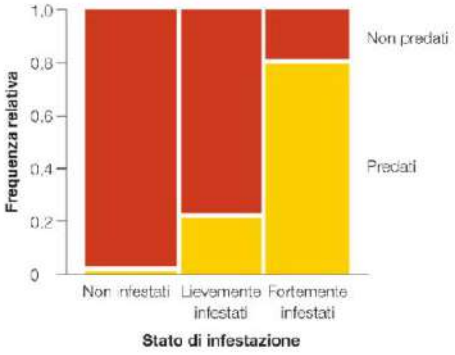


FIG. 3. A comparison of the proportion of fish estimated to have been eaten by birds after 20 d (total number missing in the open pen minus the number that escaped or died in the netted pen) showing that heavily parasitized fish were preyed on more frequently than lightly parasitized fish, which were preyed on more frequently than unparasitized fish. From left to right, percentages = 1/50, 10/45, and 37/46, respectively. Error bars are 95% confidence intervals indicating that all three groups were significantly different.



Variabili categoriche: livello di infestazione e essere/non essere mangiati sono indipendenti?

IPOTESI NULLA: essere infestati e essere mangiati sono indipendenti

IPOTESI ALTERNATIVA: essere infestati ed essere mangiati NON sono indipendenti

Come sappiamo se due eventi sono indipendenti la probabilità che si verifichino entrambi è uguale alla probabilità che si verifichi uno moltiplicata per la probabilità che si verifichi l'altro.

Quindi possiamo calcolare con la regola del prodotto le frequenze ATTESE

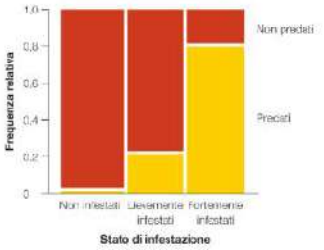
P(non infestato E mangiato)=P(non infestato)P(mangiato)

Per trovare i due fattori, basta prendere i dati in

CHI-QUADRATO

PARASSITA EUHAPLORCHIS CALIFORNIENSIS
MANIPOLAZIONE DEL COMPORTAMENTO OSPITE (PESCE)

	INFESTAZIONE		
PREDAZIONE	***assente***	***lieve***	***alto***
predati	**1**	**10**	**37**
non predati	**49**	**35**	**9**



Calcolare attese con regola prodotto che assume indipendenza (H_0)

es. Pr[*non parassitato* E *predato*] = Pr[*non parassitato*] x Pr[*predato*]

Sapendo che $df=(r-1)(c-1)$ => stimare significatività statistica $X^2 = \sum_{i=1}^{g}\sum_{j=1}^{k}\frac{(n_{ij}-E_{ij})^2}{E_{ij}}$

tabella: P(non infestato)=P(predato assente)+P(non predato assente) = (1+ 49)/(1+10+37+49+35+9)
P(mangiato) = (1+10+ 37)/(1+10+37+49+35+9)
Da qui, assumendo vera l'ipotesi nulla di indipendenza, ci aspettiamo che la probabilità sia il PRODOTTO. E quindi la frequenza ATTESA sarà QUESTA PROBABILITA' MOLTIPLICATA PER IL TOTALE DEI PESCI, cioè 17. Per la frequenza OSSERVATA basta prendere i dati
P(non infestato E mangiato)=1/(1+10+37+49+35+9), e quindi la frequenza osservata è 1.
Variabile di RISPOSTA: mangiati/non mangiati. Variabile ESPLICATIVA l'altra

abella 9.3-2

Frequenze attese di pesci predati e non predati da uccelli secondo lo stato di infestazione da trematodi.

	Non infestati	Lievemente infestati	Fortemente infestati	Totali delle righe
Mangiati dagli uccelli	17,0	15,3	15,7	48
Non mangiati dagli uccelli	33,0	29,7	30,3	93
Totali delle colonne	50	45	46	141

$$\chi^2 = \frac{(1-17{,}0)^2}{17{,}0} + \frac{(49-33{,}0)^2}{33{,}0} + \frac{(10-15{,}3)^2}{15{,}3} + \frac{(35-29{,}7)^2}{29{,}7} + \frac{(37-15{,}7)^2}{15{,}7} + \frac{(9-30{,}3)^2}{30{,}3}$$

$$= 69{,}5$$

Per i gradi di libertà, abbiamo il numero di righe (2) e colonne (3) della tabella. Il numero di gradi di libertà sarà quindi (r-1)(c-1)=2
I gradi di libertà in una tabella di contingenza (nel nostro caso 2x3) si calcolano in modo diverso, perché **non c'è il parametro da stimare**, e quindi è (r-1)(c-1) con r numero di righe e c numero di colonne. Se no dovrei anche togliere il parametro da stimare. A questo punto posso applicare la regola del chi quadro.
DEVO OGNI VOLTA SCRIVERE COSA VALE IL SINGOLO VALORE: chi quadro = … poi DL = …

Da qui, osservando la tabella per un chi quadro con due gradi di libertà, si dimostra che l'ipotesi nulla deve essere rifiutata.

FORMULA RAPIDA PER CALCOLARE LE FREQUENZE ATTESE

$$atteso[riga\ i, colonna\ j] = \frac{(totale\ riga\ i)(totale\ colonna\ j)}{totale\ generale}$$

(50x48)/141=17,0

Sono già le frequenze, non le probabilità
La dimostrazione matematica è la semplificazione di "totale generale"

CHI-QUADRATO

CALCOLO FREQUENZE ATTESE $atteso[riga\ i, colonna\ j] = \frac{(totale\ riga\ i)(totale\ colonna\ j)}{totale\ generale}$

$atteso = Pr[riga] \times Pr[colonna] \times totale\ generale$

$atteso[riga\ i, colonna\ j] = \left(\frac{totale\ riga}{totale\ generale}\right) x \left(\frac{totale\ colonna}{totale\ generale}\right) x\ totale\ generale$

ATTESO = $TOT_{[RIGA]} \times TOT_{[COLONNA]}$ / TOTALE GENERALE

GRADI DI LIBERTA' $df = (n\ righe - 1)(n\ colonne - 1)$

ASSUNZIONI <20% celle Freq attesa < 5

NO celle con Freq attesa < 1

TEST ESATTO O DI PERMUTAZIONE

P VALUE DI ASSOCIAZIONE con stesse assunzioni del chi quadro: se non sono rispettate si fa FISHER se è 2x2

TEST ESATTO DI FISHER

È il test che fornisce un P-value esatto per un test di associazione in una **tabella di contingenza 2x2**. È MIGLIORE DEL CHI QUADRO quando le assunzioni del chi quadro non ci sono e le frequenze attese delle celle sono tropo basse.

È un calcolo lungo, e si può usare solo in una tabella 2x2. È un test di PERMUTAZIONE.

La cosa che fa il test di Fisher è indicare tutte le possibili tabelle che sono estreme uguali o più estreme di quella osservata, assumendo sempre vera l'ipotesi nulla di indipendenza.

TEST ESATTO DI FISHER

PER TABELLE 2 X 2

DONNE		Trattament	(10 anni)
		SI	NO
Cancro	SI	a=2	b=4
	NO	c=8	d=1

$$p = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}$$

THM: il test esatto di Fisher rappresenta una soluzione sempre valida

Il P-value per il test di Fisher è la somma delle probabilità di tutte le tabelle estreme sotto l'ipotesi di indipendenza. Non mi confronto più con una curva teorica, ma con una curva che si ottiene dalle permutazioni, che è esattamente la stessa. **Il test di permutazione è più potente e robusto, ma a volte è molto lento computarlo.** Quindi chiedo alla macchina un numero limitato di valori che può calcolare in un tempo ragionevole.

Se gli n campionari sono bassi posso permettermi di fare un test del genere con tutte le permutazioni. La cosa che viene fatto con il test è calcolare il p value.

Questo test fu ideato da fisher per le tabelle 2x2, e poi è stato ampliato nei test di permutazione.

IL TEST ESATTO DI FISHER DA' UN P VALUE DI ASSOCIAZIONE

ESERCIZIO DEL COVID: proviamo a farlo noi

1. **Rischio Relativo (RR):**
    - **Quando usarlo:** Il rischio relativo è spesso utilizzato in studi di coorte o sperimentali con follow-up temporale, in cui è possibile definire chiaramente le esposizioni e seguire i soggetti nel tempo per osservare gli esiti.
    - **Vantaggi:** È interpretabile e può fornire informazioni dirette sulla forza della relazione tra esposizione e risultato.
    - **Limitazioni:** Non è appropriato quando gli eventi studiati sono rari.
2. **Odds Ratio (OR):**
    - **Quando usarlo:** L'odds ratio è comunemente utilizzato in studi caso-controllo, in cui si confrontano le probabilità di esposizione tra casi e controlli.
    - **Vantaggi:** È appropriato quando gli eventi studiati sono rari (pochi), ed è spesso utilizzato in analisi retrospettive.
    - **Limitazioni:** Può sovrastimare la forza della relazione quando gli eventi sono comuni. Non fornisce direttamente una stima del rischio.
3. **Test Esatto di Fisher:**
    - **Quando usarlo:** Il test esatto di Fisher è spesso utilizzato quando il **numero di osservazioni è limitato e le ipotesi del test del chi-quadro standard potrebbero non essere soddisfatte.**
    - **Vantaggi:** È robusto anche con campioni piccoli e non richiede l'assunzione di distribuzione normale.
    - **Limitazioni:** Può mancare di potenza con campioni più grandi. È particolarmente utile con dati rari o con tabelle di contingenza piccole.

In sintesi, la scelta tra RR, OR, o il test esatto di Fisher dipende dalla natura del tuo studio, dalla struttura dei dati e dagli obiettivi specifici dell'analisi. In molti casi, è consigliabile consultare uno statistico o un esperto nel campo per valutare quale misura o test sia più appropriato per il tuo scenario specifico.

- L'OR fornisce una stima puntuale della forza di associazione, mentre il test di Fisher fornisce una valutazione della significatività statistica dell'associazione.
- L'OR è spesso calcolato in modo diretto, mentre il test di Fisher valuta l'ipotesi di indipendenza nella tabella di contingenza.

In generale, l'OR può essere preferito quando si desidera ottenere una stima quantitativa della forza di associazione, mentre il test di Fisher può essere preferito quando l'obiettivo principale è valutare la significatività statistica in contesti con dati rari o tabelle di contingenza piccole

MEGLIO FARLI ENTRAMBI (OR/RR E FISHER/CHI QUADRO E POI CONFRONTARE!!!!!!!!!!!!)

ESEMPIO MODELLO PROPORZIONALE: COVID-19

FREQUENZA DI POSITIVI AL COVID-19

Gruppo	Classe d'età	Non vaccinati	Vaccinazione completa
Popolazione	12-39	3.834.795	12.185.925
	40-59	3.233.698	14.799.450
	60-79	1.345.009	11.917.105
	80	234.849	4.229.296
Diagnosi	12-39	16.805	10.123
	40-59	12.924	15.271
	60-79	4.614	11.006
	80	1.055	4.591

NON ESSERE VACCINATI AUMENTA IL RISCHIO DI INFEZIONE?

H_0: ? LA FREQUENZA DI INFETTI TRA SOGGETTI VACCINATI E NON VACCINATI E' UGUALE

H_A: ? LA FREQUENZA DI INFETTI TRA SOGGETTI VACCINATI E NON VACCINATI E' DIVERSA

COME FARESTE?

THM: alle volte (il piu' delle volte) i dati sembrano gia' pronti ma non lo sono.

ESEMPIO MODELLO PROPORZIONALE: COVID-19 - 2

IDENTIFICARE LE FREQUENZE (CELLE) DA CONFRONTARE

STIMARE FREQUENZE ATTESE

CALCOLARE GLI SCARTI PER CELLA

CALCOLARE IL CHI-QUADRATO P-VALUE

DECIDERE L'INFERENZA SULLA BASE DEL P-VALUE

THM: alle volte bisogna stimare dei parametri, e i calcoli possono essere laboriosi, ma e' semplice concettualmente.

A Massolo, Dip Biologia, Univ Pisa | alessandro.massolo@unipi.it

GRAFICI A BARRE: di solito non si usano, ma ha senso usarle quando anche lo zero è un valore possibile! Non ha senso usarla se i valori bassi non sono biologicamente sensati. Se faccio la barra mi devo assolutamente ricordare di STIMARE GLI INTERVALLI DI CONFIDENZA. Ho due metodi per stimarli (già visti)

LEZIONE 14

RELAZIONI TRA VARIABILI E STATISTICHE

	CONFRONTO		ASSOCIAZIONE
NOMINALE	BINOMIALE; $\chi 2$		$\chi 2$
ORDINALE	$\chi 2$	?	?
CONTEGGI	(BINOMIALE; POISSON)+$\chi 2$	? II PARTE	? III PARTE
NUMERICA	IC	?	?

THM: *different aims, different tools*

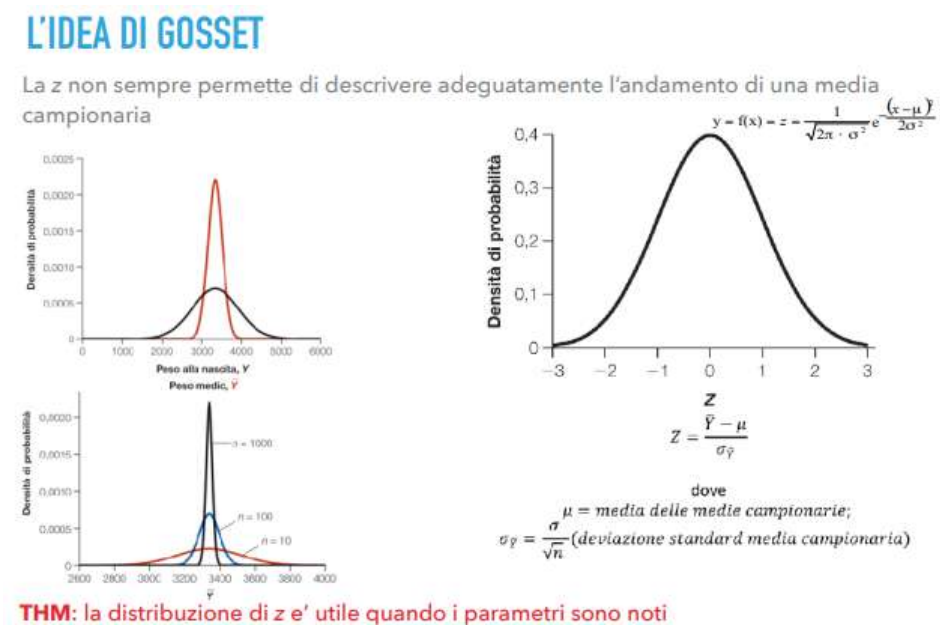
TEST T DI STUDENT

Parliamo di **distribuzione campionaria**, cioè la **DISTRIBUZIONE DI PROBABILITA' DELLA MEDIA CAMPIONARIA** (la distribuzione di probabilità di tutti i valori di una stima che possiamo ottenere campionando più volte una popolazione).

Come si fanno i confronti tra due medie

Test più usato in assoluto in biostatistica

L'idea di Gosset è legata al fatto che la Z spesso non permette di descrivere la media campionaria perché la deviazione standard è non nota, si basa su conoscenze che spesso non abbiamo. Importanti conseguenze perché al variare del n campionario ho un cambiamento delle medie campionarie. La mia stima di mu e sigma è solida se ho n alto. Se invece ho n basso la Z soffre. Quello che Gosset disse è che non sa le distribuzioni reali, e vuole il confronto tra le medie. Quindi fa la **STATISTICA t (che cambia al cambiare dei gradi di libertà) CON N ELEVATO APPROSSIMA AD UNA NORMALE.**



THM: la distribuzione di z e' utile quando i parametri sono noti

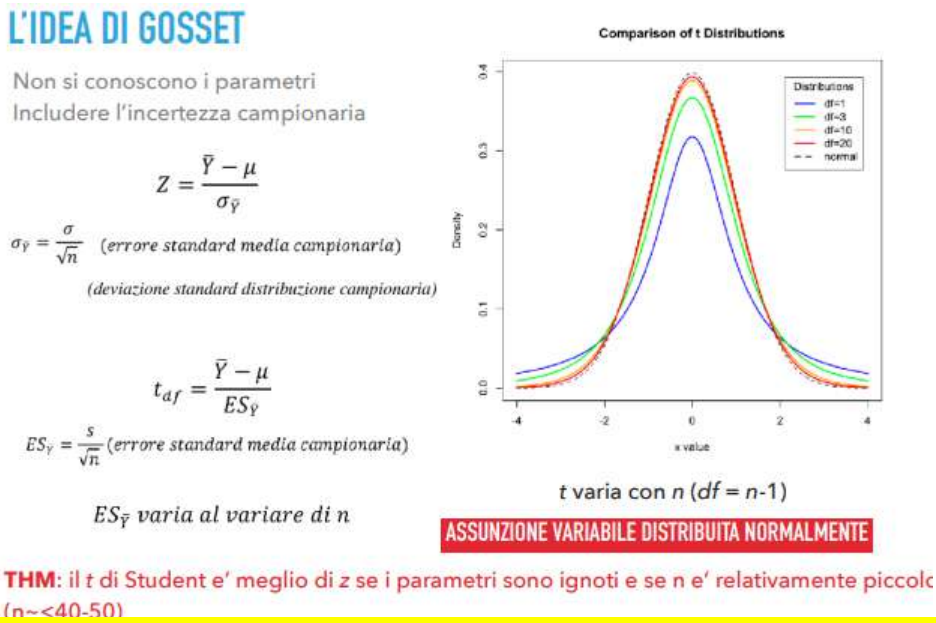
La distribuzione delle medie campionarie è una **distribuzione NORMALE** se la variabile di cui si calcola la media campionaria è anche lei distribuita normalmente. Possiamo quindi **usare la variabile standardizzata Z per calcolare le probabilità della distribuzione campionaria usando le aree sottese dalla curva normale**.

$$Z = \frac{\bar{Y} - \mu}{\sigma_{\bar{Y}}}$$

È spesso però molto **difficile calcolare Z**, perché si deve conoscere sigma, che non è quasi mai noto. Si può usare però la STIMA dell'errore standard, per cui il numeratore è la deviazione standard campionaria, che conosciamo, mentre il denominatore è n, anche questo che conosciamo (vedi formula nella slide)
Per questo viene usata t, che non sfrutta sigma come Z ma sfrutta l'errore standard.

$$t = \frac{\bar{Y} - \mu}{ES_{\bar{Y}}}$$

La distribuzione campionaria di questa statistica NON è una normale, ma un po' diversa, e questo perché



NON viene usato sigma MA l'ERRORE STANDARD, che NON è una COSTANTE come sigma, ma una VARIABILE, e cambia per effetto del caso da campione a campione.
Quindi t è meglio di Z se i parametri sono ignoti e se n è relativamente piccolo

La distribuzione t non coincide con quella di Z perché prendo ES e non sigma. Facendo così, però, AGGIUNGO L'ERRORE DI CAMPIONAMENTO ASSOCIATO ALLA STIMA DI s. quindi ho una DISTRIBUZIONE PIU' AMPIA DI QUELLA NORMALE STANDARD.
Quindi **t cambia al variare di n**, motivo per cui t non ha una sola coda di riferimento ma un numero n-1 di code di riferimento, che sono associate a n-1 con n gradi di libertà
Come calcoliamo **i GRADI DI LIBERTA'**?
In questo caso sto parlando di **un campione solo**!
Il numero di gradi di libertà è dato dal **numero di valori indipendenti (cioè n) MENO UNO, che è dovuto al fatto che sto usando un valore stimato per costruire la statistica**, cioè ES → n-1
A mano a mano che i gradi di libertà aumentano il t approssima a Z
Una assunzione importante è che la variabile di cui si sta parlando sia DISTRIBUITA NORMALMENTE, e questo perché la **deviazione standard si calcola dalla media, ma se la media non è ben distribuita non ha senso usarla.**
Se i parametri sia mu che sigma sono ignoti, soprattutto sigma, il t funziona meglio ma quando n è piccolo. se invece n è grande (tipo sopra i 40) non è sensato usare t al posto di Z. Il senso è che in teoria se n è basso e non ho informazioni sulla deviazione standard della popolazione devo usare t.
Quando n aumenta molto le curve diventano uguali, e quindi per **n che tende a numeri elevati il valore di t critico che lascia gli estremi 2,5 da un lato e dall'altro è 1,96, cioè è uguale a Z**.
Quindi **l'IPOTESI NULLA è CHE LA MEDIA CAMPIONARIA NON DIFFERISCE DAL PARAMETRO NOTO**. QUINDI POSSO USARE Z AL POSTO DI T PERCHE' s È CIRCA SIGMA.

Es. campione casuale di 4 individui: ho 3 gradi di libertà (vedi la curva verde sopra). La distribuzione normale è quella tratteggiata. Più sono pochi i gradi di libertà (aka più piccolo è il campione), meno simile è la distribuzione t alla normale. **La differenza principale NON è tanto nel picco quanto nelle CODE**, e questo è molto importante quando si calcolano gli INTERVALLI DI CONFIDENZA. Per esempio, una t con 4 gradi di libertà ha il 95% dell'area sottesa tra -2,78 e 2,78. Nel caso della distribuzione normale standard, invece, il 95% dell'area sottesa alla curva è SEMPRE tra -1,96 e 1,96. Il più ampio intervallo di confidenza

INTERVALLI DI CONFIDENZA E TEST T DI STUDENT PER UN CAMPIONE

Valutare precisione campionaria con intervallo di confidenza

Confrontare media con una media teorica attesa

$$\bar{Y} - t_{\alpha(2),df} ES_{\bar{Y}} < \mu < \bar{Y} + t_{\alpha(2),df} ES_{\bar{Y}} \quad (IC\ media\ campionaria)$$

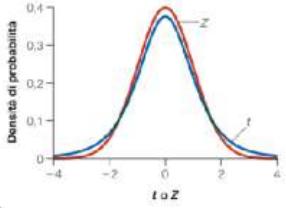
$$t_{df} = \frac{\bar{Y} - \mu}{ES_{\bar{Y}}}$$ PARAMETRO DI RIFERIMENTO

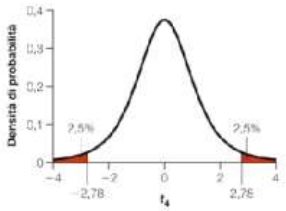
$$ES_{\bar{Y}} = \frac{s}{\sqrt{n}} (errore\ standard\ media\ campionaria)$$

$$df = n - 1$$

H_0=la media campionaria NON differisce dalla media di riferimento

H_1=la media campionaria differisce dalla media di riferimento





THM: il *t* di Student e' semplice

in t rispetto a Z è legato al fatto che non conosciamo con esattezza sigma e quindi dobbiamo stimare un errore standard, cosa che AUMENTA L'INCERTEZZA. Infatti, **la distribuzione della media campionaria che si ottiene sarà compresa entro 2,78 errori standard stimati a partire dalla media vera della popolazione**.

Come ho calcolato però 2,78? Ogni distribuzione t ha il suo valore critico al 5% che dipende dal numero di gradi di libertà. Si usa una TAVOLA STATISTICA per trovare t al 5%. Si scrive $t_{0,05(2),df}$, dove (2) indica che il valore deve essere diviso per due (aka quel valore si riferisce al 5% di area totale, diviso in 2,5% di area sottesa a destra e altrettanto a sinistra), mentre df sono i gradi di libertà e 0,05 è la percentuale totale. Da qui uso la formula dell'intervallo di confidenza scritta nella slide.

Ma perché?

Possiamo usare t per calcolare l'intervallo di confidenza per la media di una popolazione con una distribuzione normale.

Se conosciamo la MEDIA CAMPIONARIA e la DEVIAZIONE STANDARD CAMPIONARIA, vogliamo capire il GRADO DI PRECISIONE PER LA STIMA DELLA MEDIA. calcoliamo quindi l'intervallo di confidenza al 95% per il numero di gradi di libertà che abbiamo. Usiamo la distribuzione t, e questo perché sappiamo la media campionaria e parametrica e conosciamo l'errore standard, sapendo s.

Quindi nel 95% dei campioni casuali estratti la distribuzione t sarà compresa tra

$$-t_{0,05(2),df} < \frac{\bar{Y} - \mu}{ES_{\bar{Y}}} < t_{0,05(2),df}$$

Da qui si trova la formula scritta sopra, basta moltiplicare per l'errore standard e togliere la media campionaria.

Questo perché **NOI NON SAPPIAMO LA MEDIA, è LA COSA CHE VOGLIAMO STIMARE!!!**

Quindi trovo un intervallo di confidenza in cui ho alta probabilità di trovare la media. **l'idea è che questo intervallo di confidenza contiene la media della popolazione nel 95% dei campioni casuali**.

TEST t PER UN CAMPIONE

I **TEST t** sono dei metodi basati sulla distribuzione t il cui scopo è quello di verificare delle ipotesi sulle MEDIE delle popolazioni. In particolare, il **TEST t PER UN CAMPIONE viene usato per CONFRONTARE LA MEDIA DI UN SINGOLO CAMPIONE CON IL VALORE DELLA MEDIA SPECIFICATO DALL'IPOTESI NULLA**.

IPOTESI NULLA: la media vera è uguale al valore μ_0 che ipotizzo

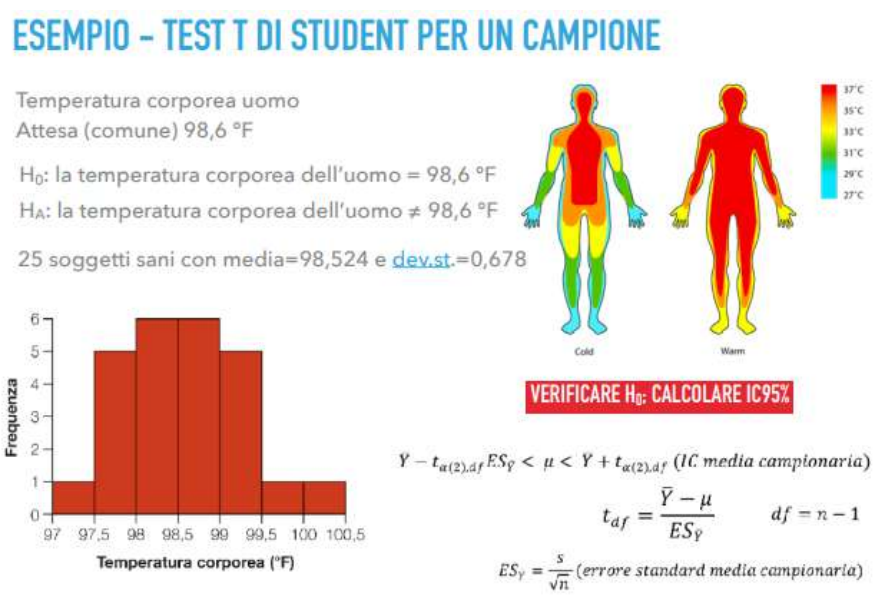
IPOTESI ALTERNATIVA: la media campionaria NON è uguale al valore μ_0

Es. se voglio dimostrare che μ_0=0, posso usare il test t per un campione, trovando la media campionaria e vedendo se è sufficientemente diversa da zero per giustificare il fatto che viene rifiutata l'ipotesi nulla, perché dalla media campionaria ho una stima di quanto è distante o meno la media vera, di cui voglio campire la distanza da μ_0

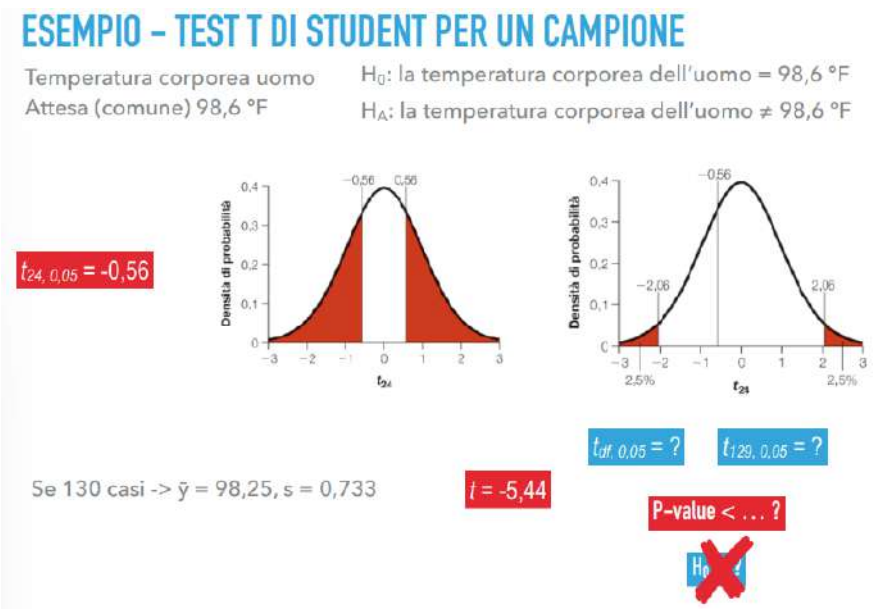
Quindi **IL TEST t per UN CAMPIONE confronta la MEDIA DI UN CAMPIONE CASUALE (AKA LA MEDIA CAMPIONARIA) CON UN VALORE DELLA MEDIA DELLA POPOLAZIONE CHE SI VORREBBE SIA GIUSTO, DEFINITO DALL'IPOTESI NULLA**.
La STATISTICA TEST è ovviamente:

$$t = \frac{\bar{Y} - \mu_0}{ES_{\bar{Y}}}$$

(Occhio a μ_0, che è la media della popolazione proposta dall'ipotesi nulla)
La distribuzione campionaria di questa statistica test, quindi, **ASSUMENDO VERA L'IPOTESI NULLA (visto che dentro alla formula ho μ_0), sarà una DISTRIBUZIONE t con n-1 gradi di libertà**. Quindi possiamo calcolare il P value confrontando la t osservata con la distribuzione t di student data l'ipotesi nulla.
Esempio: temperatura corporea dell'uomo, che si attribuisce in media 98,6 Farenheit. Abbiamo i dati relativi a 25 soggetti, che sono riportati nel diagramma a barre in questa slide. **Vogliamo sapere se i valori trovati per le temperature corporee sono in linea con la media supposta (cioè, 98,6 Farenheit) o no**.



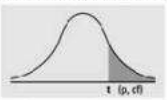
IPOTESI NULLA: la temperatura corporea dell'uomo è in media 98,6
IPOTESI ALTERNATIVA: la temperatura corporea NON è in media 98,6
Il test è A DUE CODE: sia una media troppo superiore che troppo inferiore a 98,6 Farenheit porterebbe al rifiuto dell'ipotesi nulla.
Sappiamo la temperatura media, che è **98,524**, e la deviazione standard campionaria, che è **s=0,678**. Da qui posso ricavare l'errore standard, che è 0,678/(sqrt(25))
Ora possiamo usare la statistica t per vedere di quanto la media si scosta dal valore che ci si aspetta. Se la differenza tra la media campionaria e μ_0 fosse troppo alta, la t che abbiamo osservato sarà in una delle due code con 24 (cioè, n-1) gradi di libertà.



Ora possiamo calcolare t, usando la media campionaria a cui tolgo μ_0 e divido per l'errore standard campionario.
L'ACCORDO PERFETTO TRA IPOTESI NULLA E DATI IMPLICHEREBBE CHE LA MEDIA CAMPIONARIA SIA μ_0 E t=0.
Trovo che t calcolato con μ_0 è -0,56. A questo punto posso usare la tabella per capire quanto vale t0,05.

Per df=24 trovo che t0,05 vale 2,06 (NELLA TABELLA COME SI VEDE DALL'IMMAGINE FA RIFERIMENTO ALLA SINGOLA CODA, QUINDI DEVO CERCARE IN 0,025 E NON 0,05)

Numbers in each row of the table are values on a *t*-distribution with (*df*) degrees of freedom for selectec right-tail (greater-than) probabilities (*p*).



df/p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0.259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581	4.4370
12	0.259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454	43178
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.753050	2.13145	2.60248	2.94671	4.0728
16	0.257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078	4.0150
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823	3.9651
18	0.257123	0.688364	1.330391	1.734064	2.10092	2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495
21	0.256580	0.686352	1.323188	1.720743	2.07961	2.51765	2.83136	3.8193
22	0.256432	0.685805	1.321237	1.717144	2.07387	2.50832	2.81876	3.7921
23	0.256297	0.685306	1.319460	1.713872	2.06866	2.49987	2.80734	3.7676
24	0.256173	0.684850	1.317836	1.710882	2.06390	2.49216	2.79694	3.7454
25	0.256060	0.684430	1.316345	1.708141	2.05954	2.48511	2.78744	3.7251
26	0.255955	0.684043	1.314972	1.705618	2.05553	2.47863	2.77871	3.7066
27	0.255858	0.683685	1.313703	1.703288	2.05183	2.47266	2.77068	3.6896
28	0.255768	0.683353	1.312527	1.701131	2.04841	2.46714	2.76326	3.6739
29	0.255684	0.683044	1.311434	1.699127	2.04523	2.46202	2.75639	3.6594
30	0.255605	0.682756	1.310415	1.697261	2.04227	2.45726	2.75000	3.6460
z	0.253347	0.674490	1.281552	1.644854	1.95996	2.32635	2.57583	3.2905
CI	———	———	80%	90%	95%	98%	99%	99.9%

Quindi il 5% dell'area sottesa alla curva t con df=24 si trova da una parte e dall'altra rispetto a -2,06 e 2,06. Questo vuol dire che t ricade in quell'intervallo. Quindi **NON POSSO RIFIUTARE L'IPOTESI NULLA!!!**

Quindi ad un livello di significatività di 0,05 non possiamo rifiutare l'ipotesi nulla. Ma questo vuol dire che è vero che la temperatura media è 98,6? **NON PER FORZA**!

[
Il parametro che stimiamo con il campione nel t test è la MEDIA, e **noi sappiamo che la media viene ben descritta se la variabile è distribuita in modo NORMALE, e quindi noi ASSUMIAMO UNA DISTRIBUZIONE NORMALE**.
Assunzione importante del t test: **tutti i dati devono essere INDIPENDENTI**.
Posso fare un box plot iniziale per avere una distribuzione informativa del dato. NEI COMPITI è IMPORTANTE INIZIARE CON UN BOX PLOT, che mi dà informazioni sulla simmetria della distribuzione.
La statistica t sui 25 campioni osservati (aka 24 gl) è nell'esempio un valore molto basso. La statistica t è molto simile al valore normale ma ha le code un po' rialzate.
]

Ma cosa succede se al posto di 25 persone viene preso un campione di 130 individui (con media 98,25 e s=0,733)? se calcoliamo t per df=129 troviamo che t=-5,44. Da qui si trova che l'intervallo di confidenza è tra 98,12 e 98,38, quindi la media ancora una volta sta dentro all'intervallo di confidenza. Perché abbiamo provato con un n campionario più grande? Con un campione più grande l'errore di campionamento che si commette stimando la media diminuisce, anche se il campione più piccolo aveva per effetto del caso una media stimata non molto diversa da quella del campione più grande, l'intervallo di confidenza era molto più largo.

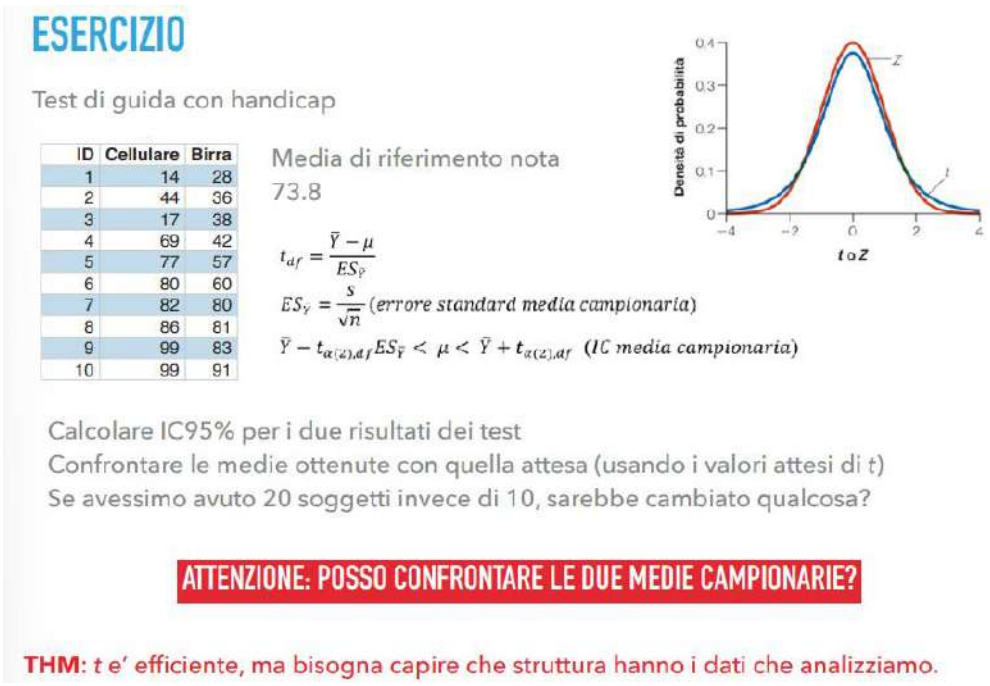
ASSUNZIONI DEL TEST t PER UN CAMPIONE

- Il campione è estratto in modo CASUALE dalla popolazione
- Ogni dato ricavato è indipendente dagli altri
- La variabile ha una distribuzione NORMALE nella popolazione

LEZIONE 14

Esercizio sulla guida della macchina

Esercizio di guida per verificare quanto un guidatore è bravo o no. Punteggio da 0 a 100. Tutti i soggetti hanno una media di riferimento di 74%. Prima si fa fare un task al cellulare e ad altri si dà a birra, e ci si chiede quali di questi due disturbi si distanzia di più dalla media di riferimento nota. **Aka do due score e vedo quale dei due si distanzia dalla media nota**. Calcolo con excel, vedendo quanto gli score si discostano dalla percentuale iniziale. Conosco l'n campionario e devo calcolare gli **errori standard**: deviazione standard diviso la radice quadrata di n. Questo lo faccio con excel anche NEL COMPITO. CON EXCEL POSSO CALCOLARE DIRETTAMENTE LA DEVIAZIONE STANDARD SENZA FARE CALCOLI.



Devo fare DUE test t al campione: devo sapere se il disturbo porta ad una prestazione diversa da quella attesa nel disturbo. Media del campione meno il parametro osservato diviso l'errore standard. Siccome il t è simmetrico

ipotesi nulla: non ci sono differenze

ipotesi alternativa: ci sono differenze: in questo caso viene A DUE CODE nel caso in cui NON ipotizzi inizialmente che guidi peggio, mentre se ipotizzo inizialmente che la guida non può essere migliore mi viene di nuovo un test a una coda.

La curva t è più "puntuta", e se si stringe al centro si alza, e questo vuol dire che le code hanno più superficie, e quindi contengono una superficie maggiore rispetto alla normale, fino a che non converge.

Non possiamo rifiutare l'ipotesi nulla se vado sopra lo 0,05. Se la significatività non emerge sono ad un n campionario ridotto, cosa che vuol dire che con un certo n campionario non ho a sufficienza dati per far emergere la frequenza che voglio. Posso ipotizzare che anziché dire la radice di 10 metto 13 soggetti, cambia solo l'errore standard ma il t cambia tantissimo! Posso infatti andare a vedere il t critico per 12 gradi di libertà, a due code perché abbiamo detto a due code. Siamo a 2,179, che poi guardo e confronto con il 2,22: RIFIUTO L'HP NULLA, perché sta OLTRE il valore soglia che mi taglia da un lato e dall'altro il 2,5% della curva.

Se avessi un p value superiore a 0,05 non avrei differenze significative.

Se vogliamo riportare questo dato, nei **MATERIALS AND METHODS** avrei potuto scrivere: ***the effect of beer on driving score*** (non scrivo: il t test era maggiore di… ma scrivo i risultati del test in materiali e metodi) ***was tested using a one sample t test***

Confidence intervals at 95% of the driving scores were calculated using the Wald method, using the z/t critical value

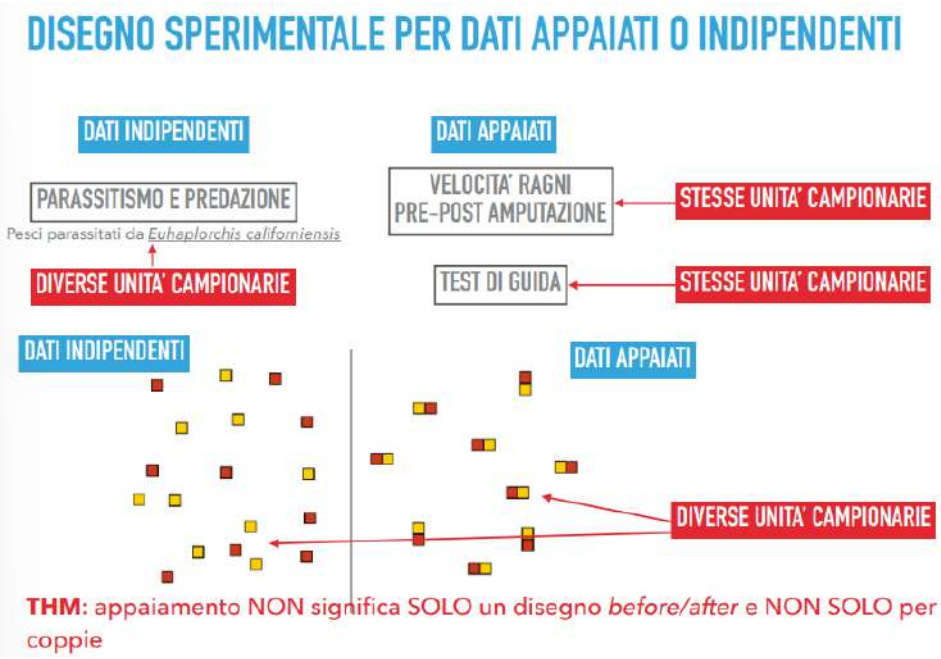
UN CAMPIONE PER LA BIRRA UNO PER I CELLULARE E FACCIO DUE t A UN CAMPIONE

QUANDO N è ALTO USO Z

QUANDO N è BASSO USO t

Da qui metto i **RESULTS**

The different treatments did not lead to significant differences OPPURE we could not detect any significant differences in driving scores in subjects whose attention was altered by cellular/beer.
Domanda: **le performance di guida delle persone alterate dal telefono rispetto a quelle alterate dalla birra cono uguali?** ora non confronto più con quello noto, ma faccio un confronto tra loro. Però i dati devono essere INDIPENDENTI. Se facessi la media dei cellulari e poi quella delle birre non sono indipendenti perché sono le stesse persone che prima fanno birra e poi fanno cellulare. Quindi non sono due gruppi indipendenti, non sono 20 dati indipendenti. L'assunto di hp nulla è che non cambia nulla se bevo birra o uso il cellulare. In questo caso mi aspetto che gli scarti siano gli stessi. Mi aspetto che ogni volta che confronto i due dati la differenza tra i due valori OGNI VOLTA tenda a zero. Quindi mi aspetto che in media le differenze siano nulle, ma IN MEDIA gli scarti tendono a zero se non c'è differenza tra le due alterazioni. **Devo confrontare la media degli scarti osservati con ZERO, che è la media attesa. Il confronto lo faccio ricalcolando con la media degli scarti t nel confronto con zero**. Uso una nuova statistica test, che è la differenza tra gli scarti in ogni persona con lo stesso ID.
La cosa che sto verificando è se le due medie sono diverse tra loro.



DATI APPAIATI O INDIPENDENTI
Dati indipendenti sono quelli che abbiamo trattato fino ad ora.
Dati invece **appaiati** sono associati al fatto che prendo due unità contigue e le confronto tra loro, ottenendo una riduzione di altri elementi di variabilità, e aumenta la capacità di disegno sperimentale.
Quindi IN UN DISEGNO AI DATI APPAIATI SI APPLICANO ENTRAMBI I TRATTAMENTI AD OGNI UNITA' CAMPIONARIA. NEL DISEGNO A DUE CAMPIONI INVECE OGNI GRUPPO DI TRATTAMENTO è FATTO DA UN CAMPIONE CASUALE DI UNITA'.
Nel disegno sperimentale dei dati appaiati però CALA n! perché **passo dall'avere n+n casi indipendenti da averne solo n, perché per ciascun dato non ho una coppia come prima**!
Nonostante ciò, i disegni a dati appaiati sono più potenti di quelli non appaiati, perché permettono di controllare molte differenze tra plot e unità campionarie a cui non siamo interessati.
Disegni appaiati anche se c'è relazione parentale, tipo con i parenti.

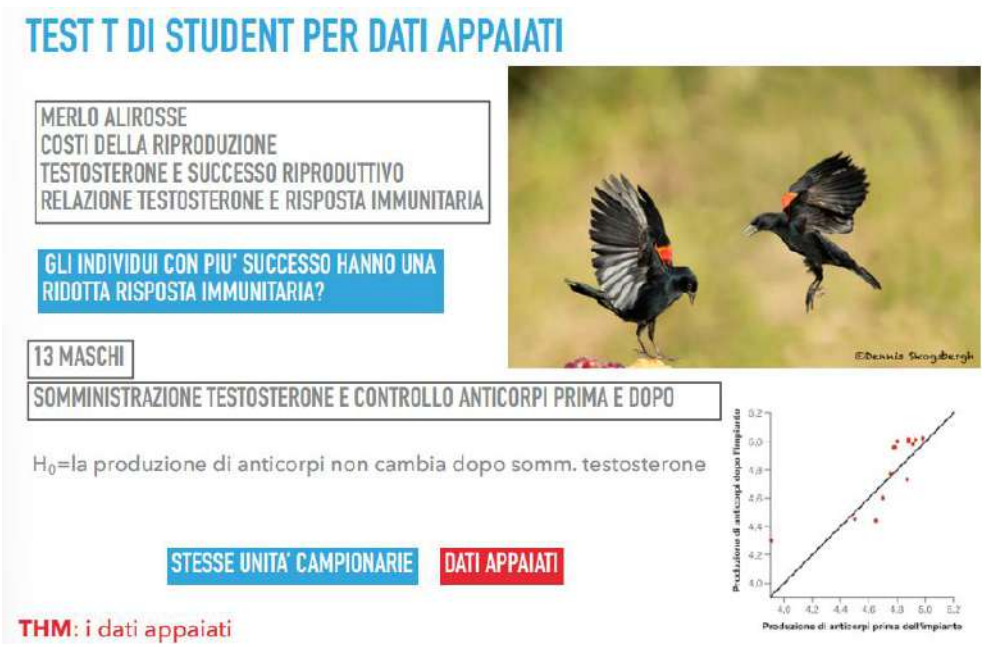
Il principale vantaggio del disegno per dati appaiati è quello di ridurre gli effetti della variabilità tra le unità campionarie che non hanno niente a che fare con il trattamento stesso. **Per esempio, se si fanno tanti plot ambientali, ciascuno avrebbe il suo campione, ma se si usa un disegno per dati appaiati si riduce l'effetto della variabilità data dall'uso di diversi campioni, visto che applichiamo i due trattamenti alle diverse metà di ogni plot**.
ESEMPIO CHE RENDE L'IDEA

- Confronto tra il peso di un soggetto prima e dopo un ricovero ospedaliero
- Confronto della diversità di specie di pesci in un lago prima e dopo una forte contaminazione da metalli

- Studio degli effetti del fumo di sigaretta in un campione di soggetti, ognuno che viene confrontato con un non fumatore della stessa età, stesso peso e appartiene allo stesso gruppo etnico. MA NON SONO LA STESSA PERSONA WTF: perché due individui diversi possono essere un dato appaiato? Lo possono essere se sono simili in virtù di caratteristiche fisiche, ambientali o genetiche.

Test t per dati appaiati

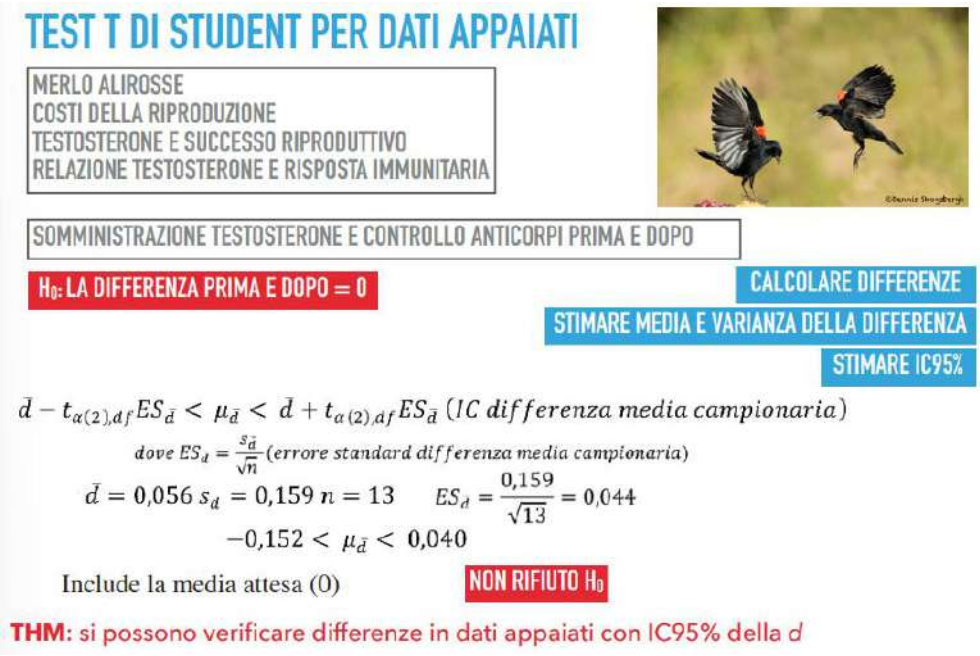
Esercizio degli uccelli: in molte specie i maschi hanno una maggiore possibilità di attrarre le femmine se presentano alti livelli di testosterone. Però i maschi con tanto testosterone potrebbero pagare questo essendo più vulnerabili alle malattie. Per verificare questa ipotesi sono stati aumentati sperimentalmente i livelli di testosterone in 13 maschi, misurando poi il livello di anticorpi prima e dopo aver aggiunto il testosterone. In questo caso parliamo quindi di **DATI APPAIATI**.



I punti vicini alla retta corrispondono alle misure e sono appunto distribuiti attorno ad una **retta inclinata di 45 gradi** (che rappresenta **NESSUN CAMBIAMENTO tra il prima e il dopo gli anticorpi**)

Mi chiedo se la differenza sia significativa. La prima cosa che si deve fare è calcolare la differenza nella produzione di anticorpi per ogni maschio (c'è una tabella), prima e dopo la somministrazione. È stata quindi presa in **considerazione una nuova variabile, d, cioè la differenza tra gli anticorpi prima e dopo**.

Ora bisogna calcolare la media campionaria di questa nuova variabile, d, e la sua deviazione standard, sulla base dei dati che abbiamo. A questo punto posso calcolare l'INTERVALLO DI CONFIDENZA, perché sapendo s so anche ES. È importante sapere che **n è il numero di COPPIE, perché l'unità campionaria è la coppia di misure** (cioè, in questo caso il singolo individuo). Quindi ho 12 gradi di libertà. Cerco il valore nella tavola e trovo che per 12 gradi di libertà t0,05=2,18. Quindi l'intervallo di confidenza è [-0,152, 0,040], che contiene lo zero come dovrebbe essere, e questo per l'ipotesi nulla (che ora vediamo)



Ora posso fare il test t normale. IL TEST t PER DATI APPAIATI SI RIDUCE A SEMPLICE TEST t DOPO CHE LE DUE MISURE PER CIASCUNA COPPIA SONO STATE RIDOTTE AD UN SINGOLO VALORE.
Quindi abbiamo

TEST T DI STUDENT PER DATI APPAIATI

MERLO ALIROSSE
COSTI DELLA RIPRODUZIONE
TESTOSTERONE E SUCCESSO RIPRODUTTIVO
RELAZIONE TESTOSTERONE E RISPOSTA IMMUNITARIA

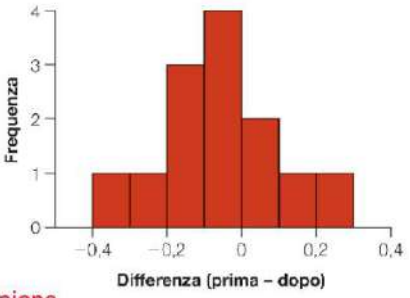
SOMMINISTRAZIONE TESTOSTERONE E CONTROLLO ANTICORPI PRIMA E DOPO

H_0=la produzione di anticorpi non cambia dopo somm. testosterone

CALCOLARE DIFFERENZE
STIMARE MEDIA E VARIANZA DELLA DIFFERENZA
STIMARE ES

$$t_{df} = \frac{\bar{d} - \mu_{d0}}{ES_{\bar{d}}}$$

dove $\bar{d}$ = *media differenza*
$ES_{\bar{d}}$ = *errore standard delle differenze*
μ_{d0} = *media delle differenze attesa sotto l'ipotesi nulla (= 0)*



THM: il test per dati appaiati e' il test *t* per un campione

IPOTESI NULLA: la variazione media della produzione di anticorpi dopo l'impianto di testosterone è zero (AKA LA **DIFFERENZA MEDIA TRA I DUE TRATTAMENTI**, quello che abbiamo calcolato prima, è ZERO)
IPOTESI ALTERNATIVA: non è zero: **TEST A DUE CODE**.
Quando calcolo t, con μ_0 che è ZERO in questo caso. Trovo quindi che
t= -1,27

TEST T DI STUDENT PER DATI APPAIATI

MERLO ALIROSSE
COSTI DELLA RIPRODUZIONE
TESTOSTERONE E SUCCESSO RIPRODUTTIVO
RELAZIONE TESTOSTERONE E RISPOSTA IMMUNITARIA

SOMMINISTRAZIONE TESTOSTERONE E CONTROLLO ANTICORPI PRIMA E DOPO

H_0=la produzione di anticorpi non cambia dopo somm. testosterone

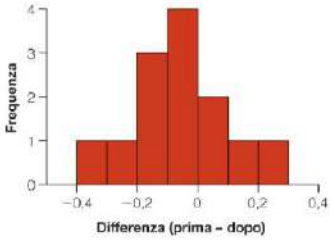
$$t_{df} = \frac{\bar{d} - \mu_{d0}}{ES_{\bar{d}}}$$

$\bar{d} = 0{,}056$ $s_d = 0{,}159$ $n = 13$ $\quad ES_{\bar{d}} = \frac{0{,}159}{\sqrt{13}} = 0{,}044$

$$t = \frac{-0{,}056 - 0}{0{,}044} = -1{,}27$$

$t_{0,05(2)(13-1)} = 2{,}18$

$t_{0,05(1)(13-1)} = 1{,}78$



NON RIFIUTO H_0

THM: il test per dati appaiati e' veloce

ho 12 gradi di libertà. **Posso calcolare il P value, MA MI SERVE UN COMPUTER**!
Quindi la cosa che faccio senza computer è usare gli intervalli di confidenza. Ho che t0,05=2,18. Quindi il mio valore è dentro l'intervallo, quindi NON POSSO RIFIUTARE L'IPOTESI NULLA.

FAI ESERCIZIO

ESERCIZIO

Amputazione pedipalpi e velocita' nei ragni genere *Tidarren*
File *spider.xls*

bef	1.25	1.64	1.91	2.31	2.37	2.38	2.84	2.87	2.93	2.94	2.98	3	3.09	3.22	3.41	3.55
afte	2.4	5.06	3.4	2.32	5.45	4.49	4.94	3.52	3.31	3.5	3.7	3.22	3.17	3.22	5.26	4.7

https://upload.wikimedia.org/wikipedia/commons/thumb/e/e8/Flickr_-_Lukjonis_-_Jumping_spiders_-_Heliophanus_sp._%28Flight%29.jpg/220px-Flickr_-_Lukjonis_-_Jumping_spiders_-_Heliophanus_sp._%28Flight%29.jpg

DEFINIRE LE IPOTESI
CALCOLARE DIFFERENZA TRA PRIMA E DOPO
CALCOLARE IC 95% DELLA DIFFERENZA TRA PRIMA E DOPO
CALCOLARE T
CALCOLARE LIVELLO DI SIGNIFICATIVITA'

TEST T DI STUDENT PER DUE GRUPPI INDIPENDENTI – VARIANZE OMOGENEE

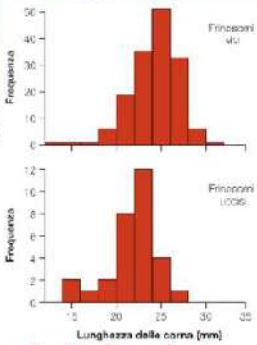
FRINOSOMI
PREDAZIONE E SPINE CUTANEE

GLI ANIMALI PREDATI HANNO SPINE PIU' PICCOLE?

Phrynosomatidae (Nord America)
Horny lizards, horny toads o horntoads

Lanius ludovicianus
Averla Stolida (loggerhead shriek)

30 ANIMALI PREDATI
154 ANIMALI VIVI

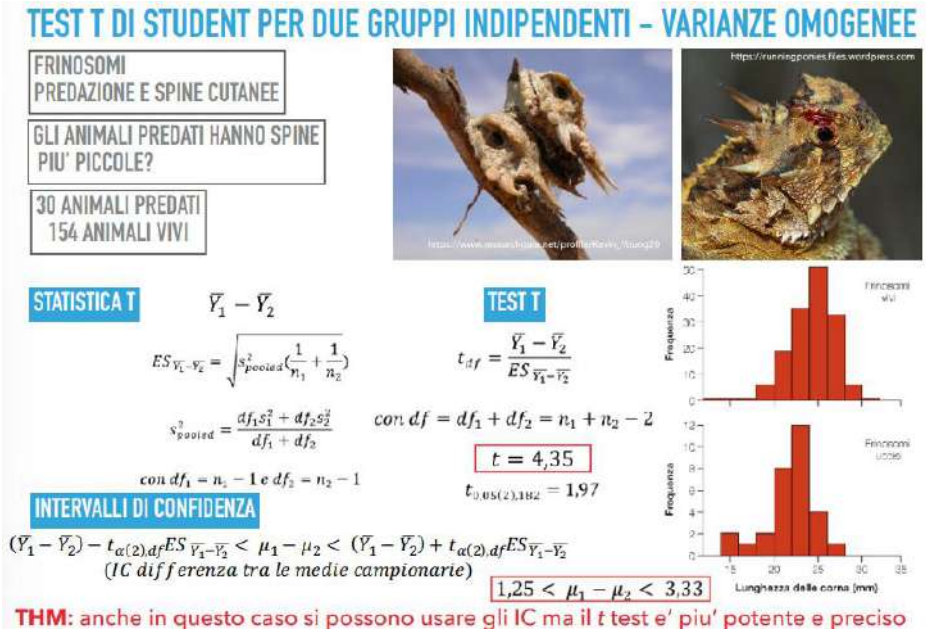


H_0: le spine degli animali predati e vivi sono ugualmente lunghe

THM: il test *t* per dati indipendenti confronta le medie di 2 campioni indipendenti

Abbiamo visto ora i dati appaiati, ma i **dati NON APPAIATI**?
CONFRONTO TRA LE MEDIE IN DUE CAMPIONI in cui i trattamenti sono applicati a due unità campionarie distinte
Esempio: lucertole. Ci si chiede se le spine cutanee hanno anche un effetto anti-predatorio nei confronti di altri organismi. La domanda è se le spine cutanee sono funzionali. Sono state confrontate le lunghezze delle spine cutanee di frinosomi che erano stati uccisi con frinosomi vivi: se le spine servissero nella difesa alla predazione, ci dovrebbe essere una differenza.
Volendo confrontare due distribuzioni è **meglio usare UN GRAFICO SOLO CON DUE BOX PLOT, l'istogramma NON dà una capacità discriminatoria elevata**. Sembra emergere una maggiore lunghezza delle spine dei soggetti vivi, piuttosto che di quelli uccisi.

La cosa che faccio è usare una statistica t diversa che si basa sulla DIFFERENZA TRA LE DUE MEDIE. Cioè, voglio capire di quanto sono più lunghe in media le corna dei frinosomi vivi. La migliore stima della differenza tra due medie è la **DIFFERENZA TRA LE MEDIE CAMPIONARIE**. $\bar{Y}1-\bar{Y}2$, con $\bar{Y}1$ per indicare la media campionaria di quelli vivi e $\bar{Y}2$ per la media campionaria di quelli morti.
Ora **calcolo gli INTERVALLI DI CONFIDENZA**: se la variabile ha una distribuzione normale in entrambe le popolazioni, la distribuzione campionaria della differenza tra le medie campionarie è una NORMALE, questo perché la distribuzione delle medie campionarie singole è normale, e data additività lo è anche la somma. Quindi posso usare la distribuzione t di student.
DEVO CALCOLARE L'ERRORE STANDARD, che è l'errore standard della differenza, di cui ho la formula, che considera le varianze dei singoli casi. In questo errore standard compare la grandezza



s^2 che è la VARIANZA CAMPIONARIA COMUNE, cioè una **MEDIA PONDERATA (sulla base dei gradi di libertà) della varianze campionarie nei due gruppi**. Si assume per l'intervallo di confidenza che le due varianze delle popolazioni siano le stesse (AKA VARIANZE OMOGENEE).

Poiché la distribuzione campionaria di $\bar{Y}_1 - \bar{Y}_2$ è normale, la distribuzione campionaria della seguente differenza standardizzata ha una distribuzione t di Student:

$$t = \frac{(\bar{Y}_1 - \bar{Y}_2)}{ES_{\bar{Y}_1-\bar{Y}_2}}$$

con un totale dei gradi di libertà uguale a

$$df = df_1 + df_2 = n_1 + n_2 - 2$$

ALTRO MODO, NON TEST T DIRETTO: posso calcolare l'intervallo di confidenza della differenza tra le due medie delle popolazioni (ultima formula in basso). Ho che t0,05 viene 1,97. Una volta trovato ES con la formula, e sapendo dai dati Ȳ1-Ȳ2 posso calcolare l'intervallo di confidenza in cui ci sta la differenza tra le due mu → $1{,}25 < \mu_1 - \mu_2 < 3{,}33$

Questo ci porta a dire che la differenza tra le due medie campionarie sta in un intervallo ABBASTANZA LONTANO DA ZERO per dire che IN EFFETTI C'E' DIFFERENZA (L'INTERVALLO NON CONTIENE LO ZERO!).

IL T TEST è LA SCELTA MIGLIORE RISPETTO AGLI INTERVALLI DI CONFIDENZA! Questo perché comunque si devono calcolare gli errori standard.

ASSUNZIONI DI QUESTO TEST

- Ogni campione è un campione casuale estratto dalla sua popolazione
- La variabile numerica in ogni popolazione ha una distribuzione normale
- **La deviazione standard e la varianza della variabile numerica è la OMOGENEA in entrambe le popolazioni**

L'ULTIMA ASSUNZIONE è NUOVA RISPETTO AL TEST t PER UN CAMPIONE.

NECESSITA' DI UN TEST COMPLICATO per le variabili disomogenee: per capire se sono disomogenee c'è bisogno di alcune assunzioni

TEST T DI STUDENT PER DUE GRUPPI INDIPENDENTI – VARIANZE NON OMOGENEE

SALMERINI DI FONTE E SALMONI REALE
12 FIUMI

I FIUMI CON I SALMERINI HANNO MENO SALMONI

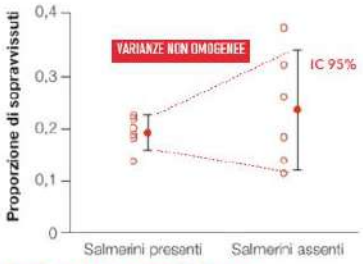
RILASCIATI SALMONI IN FIUMI CON E SENZA SALMERINI

CONTA DI SALMONI DOPO 12 MESI

12 UNITA' CAMPIONARIE (FIUMI)

$$df = \frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\left[\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1-1}+\frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2-1}\right]} \qquad t = \frac{(\bar{Y}_1-\bar{Y}_2)-(\mu_1-\mu_2)_0}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$$

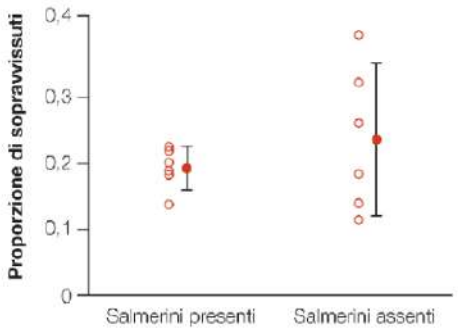
P=0,40

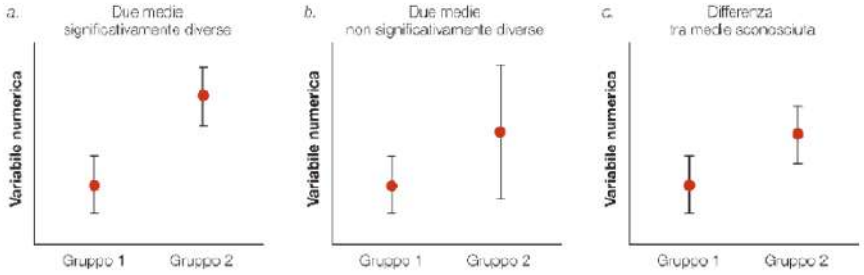


THM: con campioni con varianze disomogenee si necessita un *t* test modificato

Ma cosa succederebbe nel caso in cui le varianze non fossero omogenee? Si deve usare il test t approssimato di Welch, che confronta le medie di due gruppi e può essere usato anche quando le varianze dei due gruppi sono diverse.

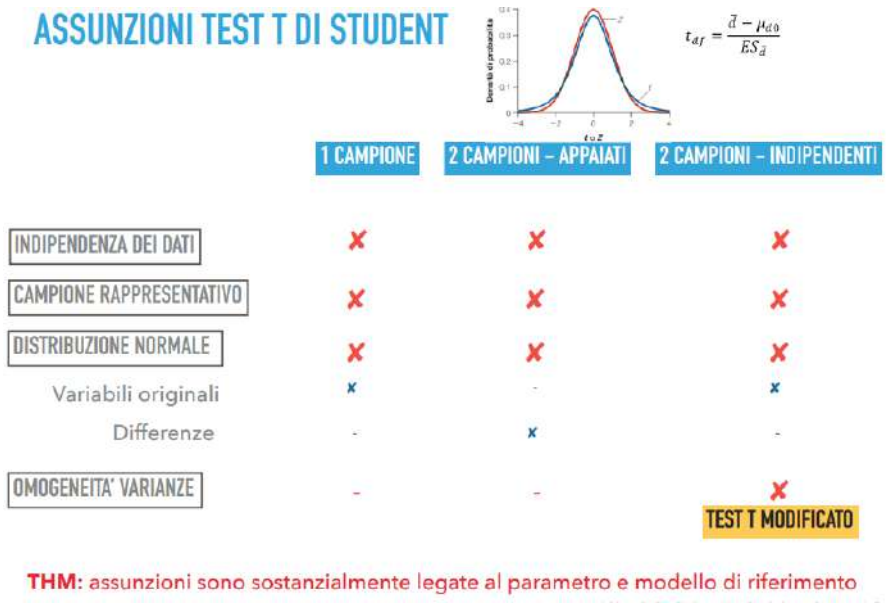
INTERPRETAZIONE IC





THM: usare cautela nel confrontare IC

Interpretazione intervalli di confidenza: quando gli intervalli non si sovrappongono per nulla (come nella figura a), l'ipotesi nulla di MEDIE UGUALI è ovviamente da rifiutare, perché NON CI SONO VALORI COMUNI. Se invece uno è compreso nell'altro (immagine b) l'ipotesi è da NON rifiutare, mentre come in c non ho informazioni certe



Quindi ho TRE T TEST (che i realtà sono due: quello a un campione e quello a dati appaiati sono gli stessi): **l'INDIPENDENZA DEI DATI è INDISPENSABILE PER TUTTI E TRE, come che il campione sia RAPPRESENTATIVO** (se no il nostro lavoro è sbagliato). Poi tutti i t test si basano sulla MEDIA, quindi anche qui ho bisogno che tutti e tre siano a distribuzione NORMALE. **Nel test a due campioni ho la DIFFERENZA però che deve essere distribuita in modo normale, mentre negli altri due mi serve che siano le variabili vere e originali ad essere distribuite in modo normale!**!

L'omogeneità delle varianze si assume nel test a due campioni indipendenti, negli altri casi no.

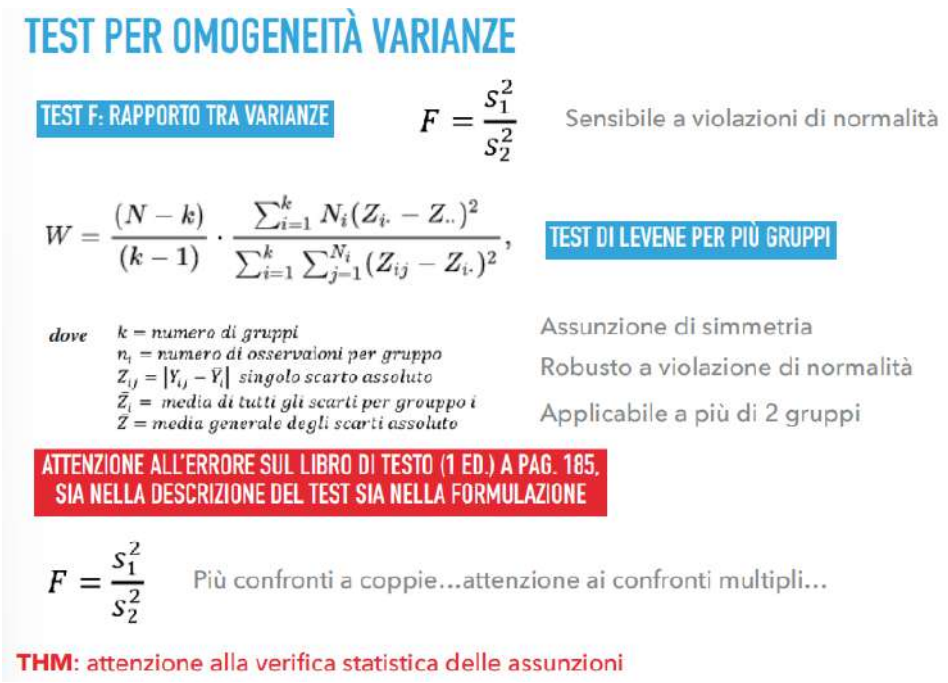
Le assunzioni sono legate al parametri del modello di riferimento

Per fare la **verifica dell'omogeneità faccio il TEST F** e valuto il rapporto (ci metto pochissimo perché i dati li ho già)

LEZIONE 16

TEST F PER OMOGENEITA' DI VARIANZE

È spesso importante capire se due popolazioni differiscono per dispersione nelle misure, non solo per le medie. Questo, per esempio, è utile per capire se si può fare un test t a due campioni.



Se ho 3 campioni indipendenti, normali ed omogenei, faccio il t test per GRUPPI INDIPENDENTI, per a con b, a con c, b con c. La cosa che però non viene presa in considerazione è che **FACCIO SOLO DUE CONFRONTI, e quindi c'è UN CONFRONTO CHE VIENE RIPETUTO**. Se io avessi **invece 4 individui e facessi due confronti non avrei dati ripetuti, mentre se ne ho 3 confronto dati più volte gli stessi**. **Se confronto più volte gli stessi dati AUMENTO LA PROBABILITA' DI ERRORE**. Quindi la cosa migliore sarebbe avere a,b,c,d e confrontare a con b, c con d.

Se invece ho tre dati la cosa che posso fare faccio i CONFRONTI MULTIPLI, cosa che mi fa ABBASSARE LA SOGLIA DI RIFERIMENTO, in modo da rendere più difficile IL RIFIUTO DELL'IPOTESI NULLA.

Il test F è molto sensibile agli scostamenti dall'assunzione di normalità, motivo per cui non è consigliabile usarlo in molti casi, perché spesso le distribuzioni NON SONO PERFETTAMENTE NORMALI.

QUINDI ENTRAMBI I CAMPIONI DEVONO ESSERE CASUALI E LA VARIABILE NUMERICA DEVE AVERE UNA DISTRIBUZIONE NORMALE IN ENTRAMBE LE POPOLAZIONI.

Se l'ipotesi nulla fosse vera, allora F dovrebbe essere vicino a 1, discostandosi da 1 solo per effetto del caso. Se è vera l'ipotesi nulla, inoltre, la statistica F ha una distribuzione F con gradi di libertà n1-1 (gradi di libertà del numeratore) e n2-1 (gradi di libertà del denominatore).

Anche il test F ha una sua tabella di valori in base ai gradi di libertà, che visto che i gruppi sono 2 il test F è formato da due set di gradi di libertà, in quanto tiene conto del campione 1 e 2 e della sua molteplicità.

Ci sono anche dei metodi alternativi al test F per verificare l'ipotesi nulla. Uno dei migliori è il **test di Levene** (vedi slide).

TEST NORMALITA'

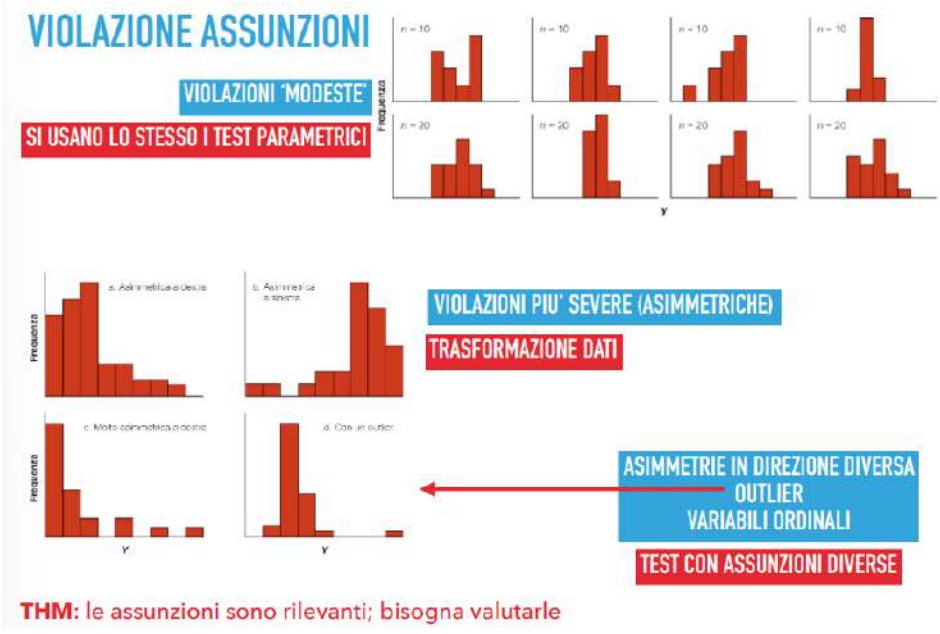
Come faccio a capire se un set di dati numerici sia estratto da una popolazione che ha una distribuzione normale o no?

H0=è NORMALE

Si usa un test detto **test di normalità**.

Con N grande i test di normalità dicono che tutte le curve sono diverse dalla normale. Quindi vale solo con un N campionario tra 40 e 50.

I metodi più semplici per capire come è fatta una distribuzione sono quelli GRAFICI. Si inizia facendo un istogramma, che spesso non rappresenta una curva perfetta, anche perché la dimensione del campione è ridotta, neanche se la popolazione di provenienza ha una distribuzione normale.



Per esempio, nella figura in alto a destra si vedono due gruppi di istogrammi relativi a dati campionati da una distribuzione NORMALE. **Nessuno di questi istogrammi assomiglia ad una gaussiana, ma nessuno ha un comportamento così anomalo da farci pensare che l'assunzione di normalità sia sbagliata**. Se invece la distribuzione di frequenza è asimmetrica o bimodale, o se ci sono outlier, è improbabile che la popolazione di partenza abbia una distribuzione normale (immagine in basso a sinistra nella slide, l'ultimo istogramma per esempio ha un outlier, gli altri tre hanno distribuzioni fortemente asimmetriche).

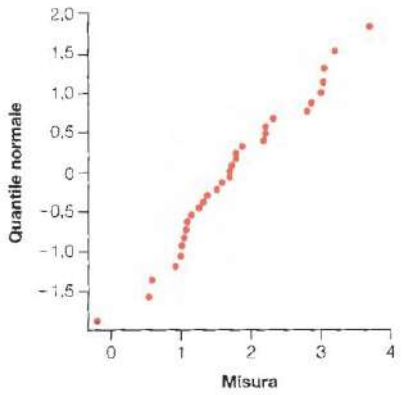


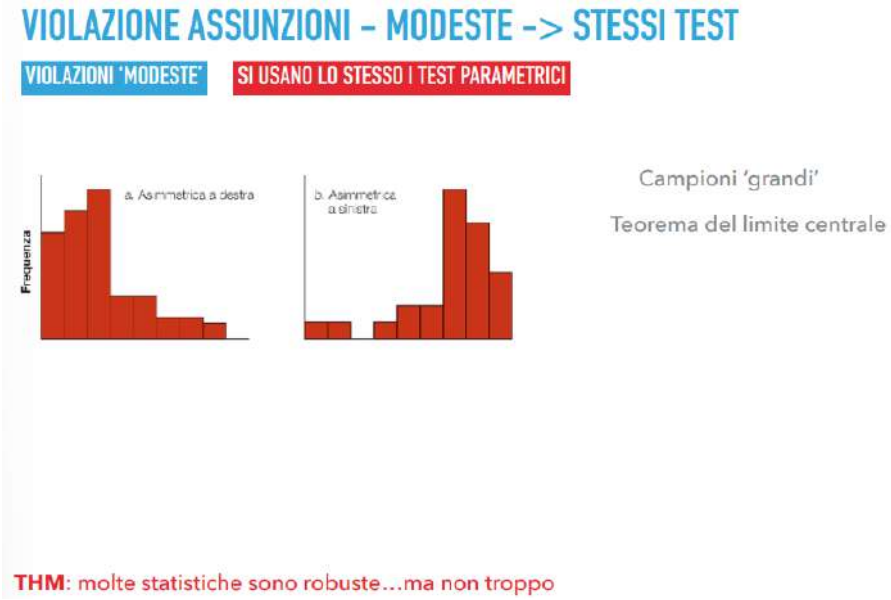
Figura 13.1-3
Un diagramma dei quantili normali per un campione casuale di 32 osservazioni estratte da una distribuzione normale. I punti del diagramma giacciono all'incirca su una retta.

Un modo per capire se una distribuzione è normale è anche usare il **DIAGRAMMA DEI QUANTILI NORMALI**: si confronta ciascuna osservazione del campione con il quantile corrispondente previsto dalla distribuzione normale standard. **se i dati sono presi da una distribuzione normale, i punti sono su una LINEA RETTA**. Ovviamente non saranno mai perfettamente su una linea retta, ma anche nella migliore delle ipotesi oscilleranno.

Come dicevo, spesso con gli istogrammi basta il COLPO D'OCCHIO.
A volte però il colpo d'occhio non basta, esistono dei test veri e propri. Per esempio:
IPOTESI NULLA: i dati sono campionati da una popolazione che ha una distribuzione normale.
IPOTESI ALTERNATIVA: i dati sono campionati da una popolazione che non ha una distribuzione normale.
Bisogna però **fare ATTENZIONE a questi test**: da un lato, il test può dare un falso senso di sicurezza. Una dimensione campionaria piccola non permette una potenza tale da rifiutare l'ipotesi nulla di normalità anche se i dati sono estratti da una popolazione normale.

Ma quando è possibile **IGNORARE LE VIOLAZIONI DELLE ASSUNZIONI**? Come dobbiamo comportarci quando i dati non soddisfano le assunzioni di normalità e di uguaglianza delle deviazioni standard?
Dipende ovviamente dal TIPO di violazioni. Possiamo in alcuni casi ignorarle, in virtù del fatto che i metodi per stimare e valutare le medie non sono molto sensibili alle assunzioni di normalità e di uguaglianza delle deviazioni standard.
Se per esempio le violazioni sono modeste, continuano ad essere usati i test parametrici che conosciamo (t, normale)

UNA PROCEDURA STATISTICA è ROBUSTA SE LA RISPOSTA CHE FORNISCE NON è SENSIBILE A VIOLAZIONI DELLE ASSUNZIONI DEL METODO.

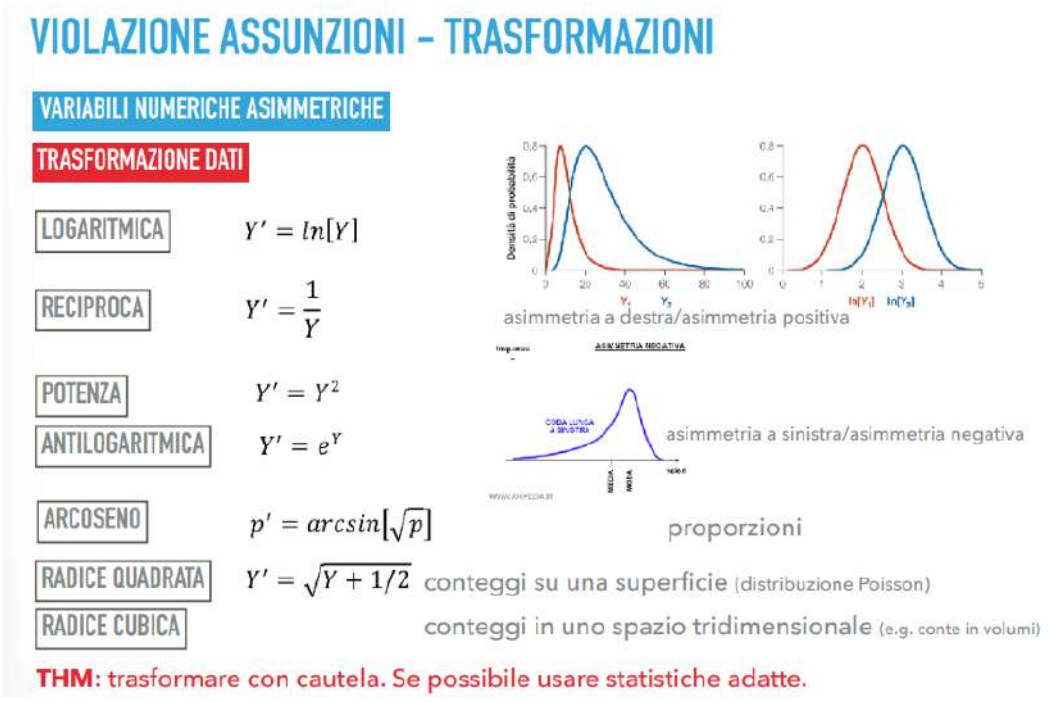


Per il calcolo degli intervalli di confidenza e del test t è sempre preferibile che i dati siano distribuiti normalmente. A volte è però possibile usare questi test anche se i dati non sono distribuiti in modo perfettamente normale. Il motivo è il TEOREMA DEL LIMITE CENTRALE! Che è ROBUSTO, e dice che la distribuzione delle medie campionarie risulta sempre di più approssimabile ad una distribuzione normale quanto più aumenta la dimensione del

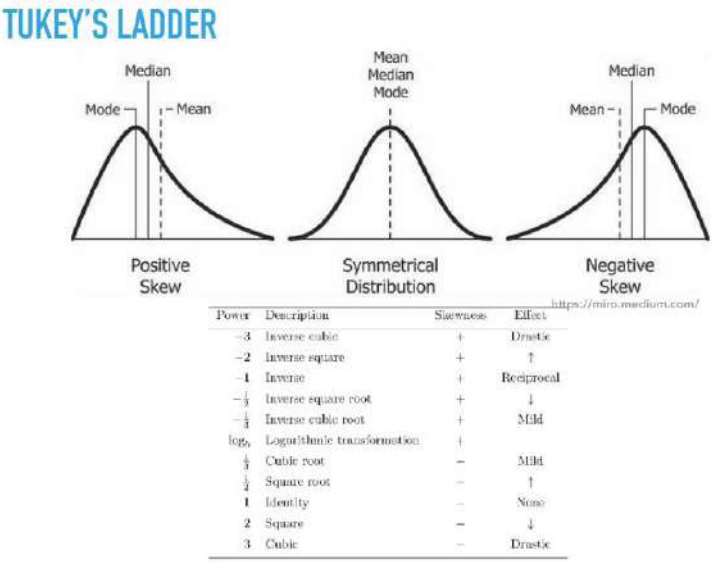
campione. Tutto questo vale sempre se la violazione della normalità non è troppo forte, se no servono quasi infiniti individui nel campione per renderla una distribuzione normale. Quindi posso fare il test t. questa robustezza vale per i test che testano le medie, **MA NON PER I TEST SULLE VARIANZE, COME IL TEST F, per cui SERVE CHE LA DISTRIBUZIONE SIA ESATTAMENTE NORMALE**.

Ma quanto devono essere grandi i campioni per poter ignorare l'assunzione di normalità?
DIPENDE DALLA FORMA DELLA DISTRIBUZIONE: se per esempio si prendono in considerazione due distribuzioni con la stessa asimmetria, si ottengono risultati accurati con un numero esiguo di unità per campione (tipo 30). Ma se e distribuzioni sono asimmetriche da due parti opposte servirebbero magari 500 unità. Le distribuzioni di frequenza che contengono gli outlier, invece, non dovrebbero mai essere analizzate con un test t. I METODI CHE ASSUMONO NORMALITA' SONO MOLTO SENSIBILI AGLI OUTLIER.

Violazioni MODESTE si usano gli STESSI TEST, quindi lo stesso i test parametrici.
Quando i dati non sono conformi alle assunzioni di un metodo statistico si sottopongono ad una TRASFORMAZIONE che li renda capaci di soddisfare meglio queste assunzioni. In questo modo posso migliorare l'adattamento della distribuzione normale ai dati e per rendere le deviazioni standard più simili in diversi gruppi. Per esempio, spesso un campione di dati non segue una distribuzione normale, ma i logaritmi sì. Se faccio quindi la procedura di stima con il logaritmo ricavo delle conclusioni esatte e sensate.



Facendo una trasformazione **CAMBIO OGNI MISURA CON LA STESSA FORMULA MATEMATICA**.
Sono due campioni, e la cosa che facciamo è fare una logaritmica, che amplifica la differenza dei valori bassi Quando ho **la asimmetria a sinistra la cosa che devo fare è stirare l'altra parte, cosa che faccio con la POTENZA**. Si allargano quindi le differenze più di quanto si allarghino le altre.
Ci sono poi anche altre trasformazioni, come arcoseno e radice quadrata, per esempio quando si fanno conteggi su una superficie. Quando invece si fanno conteggi su un volume si fa una radice cubica.



Power	Description	Skewness	Effect
-3	Inverse cubic	+	Drastic
-2	Inverse square	+	↑
-1	Inverse	+	Reciprocal
$-\frac{1}{2}$	Inverse square root	+	↓
$-\frac{1}{3}$	Inverse cubic root	+	Mild
$\log_e$	Logarithmic transformation	+	
$\frac{1}{3}$	Cubic root	−	Mild
$\frac{1}{2}$	Square root	−	↑
1	Identity	−	None
2	Square	−	↓
3	Cubic	−	Drastic

TUKEY'S LADDER OF POWER
È un **set di trasformazioni** applicate ai dati in modo da rendere le varianze più facili da analizzare. L'idea è scegliere una trasformazione che rende il dato simmetrico, per esempio tramite la trasformazione logaritmica, o radice quadrata.

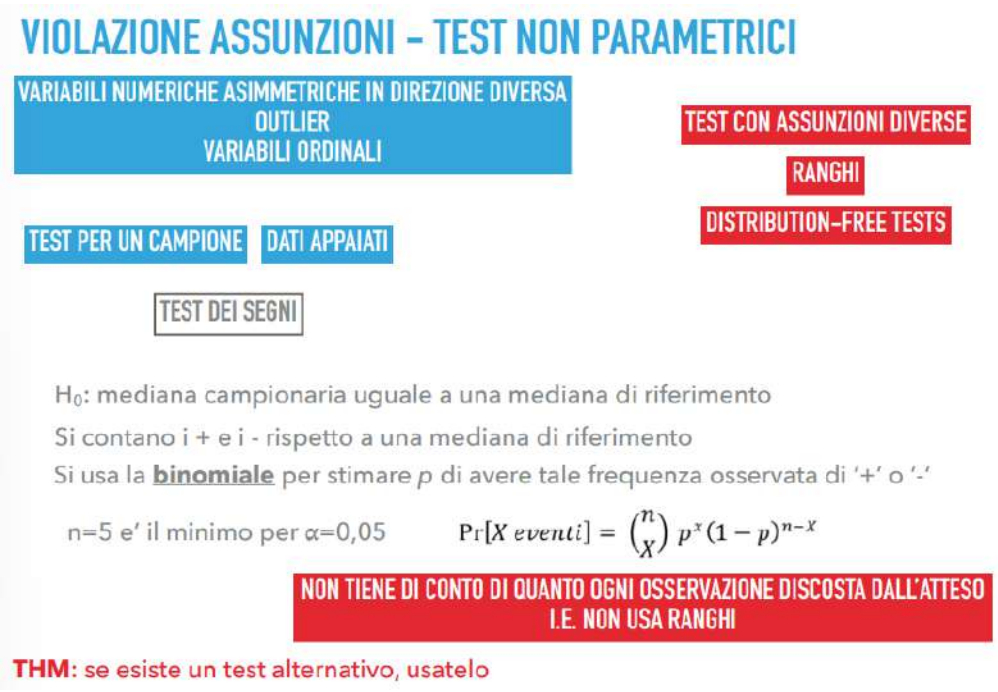
TEST NON PARAMETRICI

Se ignorare le assunzioni o trasformare i dati non basta, si possono usare TEST NON PARAMETRICI. Questi metodi necessitano di assunzioni meno restrittive riguardo alla distribuzione di probabilità della variabile di interesse. Sono quindi utili quando la **distribuzione di frequenza NON è normale, e per esempio quando ci sono OUTLIER**.

UN METODO NON PARAMETRICO FORMULA MENO ASSUNZIONI SULLA VARIABILE DI UN METODO PARAMETRICO.

ALTERNATIVE AL TEST t PER UN CAMPIONE E QUELLO PER DATI APPAIATI.

I test NON parametrici si basano di solito sui **RANGHI dei dati** (cioè, la posizione relativa del dato all'interno di un insieme **ordinato** di dati, aka mediana, quartili… oppure semplicemente la loro posizione rispetto alla lista ordinata, quindi primo o secondo o terzo) e non si loro valori effettivi. **QUINDI NEL TEST NON VENGONO USATI I VALORI MISURATI**! Ma se l'ipotesi nulla è vera, i ranghi saranno distribuiti circa nello stesso modo nei due gruppi (es. distribuzione dei valori rispetto alla mediana, o cose del genere): **se uno dei due gruppi presenta invece la maggior parte delle misure piccole, l'altro la maggior parte delle misure grandi, vuol dire che siamo di fronte alla necessità di rifiutare l'ipotesi nulla**.

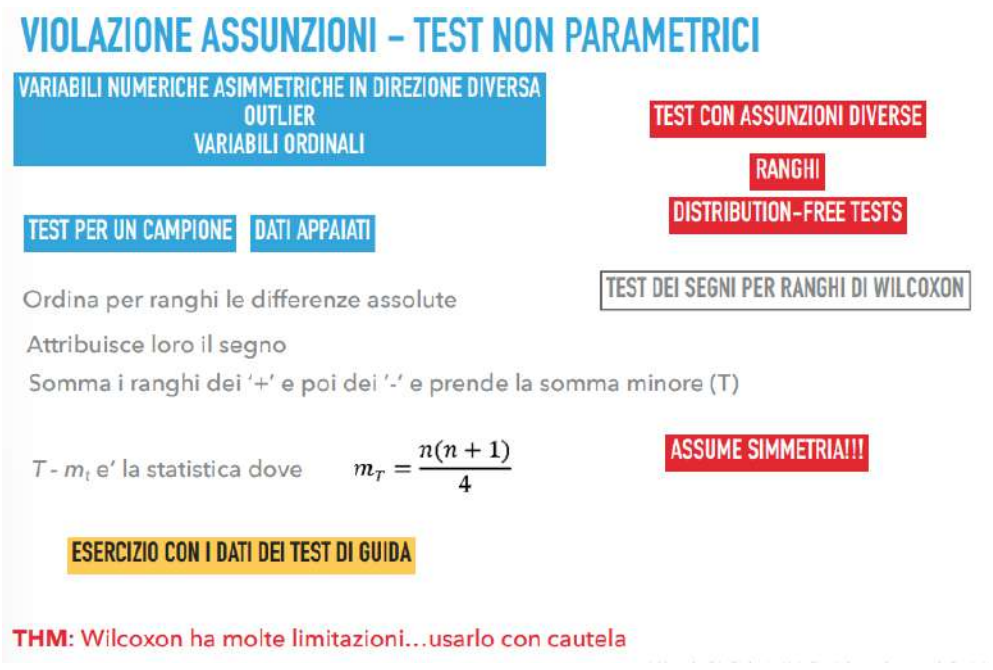


Quindi questi test sono molto utili quando ci sono outlier, che non influenzano troppo i ranghi.

Uno dei test non parametrici più usati è il **TEST DEI SEGNI**. Questo test si può usare in alternativa al test t per un campione o al test t per dati appaiati quando NON è soddisfatta l'assunzione di normalità.
Questo test valuta se la MEDIANA di una popolazione è uguale a un valore specificato dall'ipotesi nulla. Le misure che sono al di sopra del valore specificato dall'ipotesi nulla sono indicate con +, quelle al di sotto con -. **Se l'ipotesi nulla è vera, ci attendiamo che META' delle misure sia al di sopra della mediana specificata dall'ipotesi nulla e metà al di sotto**. Il test dei segni è un semplice test BINOMIALE in cui il numero delle unità maggiori o minori della mediana specificata dall'ipotesi nulla viene confrontato con quello atteso. P=0,5 (aka la P di trovare un +)
Purtroppo ha una potenza molto bassa rispetto al test t perché NON tiene conto delle informazioni contenute nei dati: per esempio, una misura solo leggermente maggiore del valore della mediana specificato dall'ipotesi nulla o un valore che lo supera molto avranno la STESSA IMPORTANZA IN QUESTO TEST! È comunque utile soprattutto in caso non si possa usare nessun altro test.

Come funziona questo test? Ci aspettiamo dall'ipotesi nulla metà dei dati sotto alla mediana e metà sopra. Ovviamente i risultati non daranno la metà esatta. Quindi dobbiamo calcolare il P VALUE?
Per esempio, qual è la probabilità di ottenere 7 o meno osservazioni + su 25 quando la probabilità di una osservazione + è il 50%?
È una **BINOMIALE**! Perché posso ottenere o – o +, quindi è un successo/insuccesso. 2P(x<=7)=2bin(7)=0,04
Questo valore è minore di 0,05, quindi si RIFIUTA l'ipotesi nulla.
Se la dimensione del campione è minore di 5 non si può rifiutare l'ipotesi in base a un test dei segni bilaterale.

Altro test dei segni che viene usato è il **TEST DEI SEGNI PER RANGHI DI WILCOXON** (vedi nella slide)
È un miglioramento rispetto al test appena visto: sono CONSERVATE LE IPOTESI SUI VALORI ASSOLUTI; quindi, viene considerato se un valore si trova tanto sopra o poco sopra alla **mediana. Purtroppo, però assume che la distribuzione delle misure nella popolazione sia SIMMETRICA attorno alla MEDIANA, cosa che è ALTAMENTE RESTRITTIVA**, e di nuovo non può considerare gli outliers. Il principale motivo per cui i dati non possono essere interpretati con una normale o con una trasformazione è proprio l'asimmetria!



Un altro test non parametrico che viene utilizzato è il **TEST U DI MANN-WHITNEY**. Può essere usato al SOLO posto del test t a due campioni quando l'ipotesi di normalità non è soddisfatta. Considera di nuovo i RANGHI delle misure per valutare se le distribuzioni di frequenza di due gruppi siano o meno uguali. non richiede tante assunzioni quante quelle del test t a due campioni, e SE DUE **DISTRIBUZIONI HANNO LA**

VIOLAZIONE ASSUNZIONI – TEST NON PARAMETRICI – MW

ESEMPIO

CANNIBALISMO SESSUALE NEI GRILLI
MASCHI OFFRONO LE ALI POSTERIORI A FEMMINE

FEMMINE AFFAMATE SONO PIU' PRONE ALL'ACCOPPIAMENTO?

PRESENTAZIONE MASCHIO A 24 FEMMINE
11 AFFAMATE E 13 NUTRITE
TEMPO ALL'ACCOPPIAMENTO

Feeding	Time	Feeding	Time
starved	1.9	fed	1.5
starved	2.1	fed	1.7
starved	3.8	fed	2.4
starved	9	fed	3.6
starved	9.6	fed	5.7
starved	13	fed	22.6
starved	14.7	fed	22.8
starved	17.9	fed	39
starved	21.7	fed	54.4
starved	29	fed	72.1
starved	72.3	fed	73.6
		fed	79.5
		fed	88.9

STESSA FORMA si può anche usare per **CONFRONTARE LA POSIZIONE DELLE MISURE DI TENDENZA CENTRALE**.

Esempio: grillo che si accoppia perché il maschio offre alla femmina le sue ali come cibo. un maschio con ali leggermente mangiucchiate ha meno probabilità di accoppiarsi. Ma quindi le femmine si accoppiano di più se sono affamate? Si risponde prendendo 24 femmine in 2 gruppi: 11 femmine a digiuno, 13 no. tutte le femmine sono poi state messe in gabbia separate, ciascuna con un maschio vergine, ed è stato registrato il tempo necessario per accoppiarsi.

IPOTESI NULLA: il tempo di attesa fino all'accoppiamento è lo stesso per le femmine tenute a digiuno e per quelle alimentate

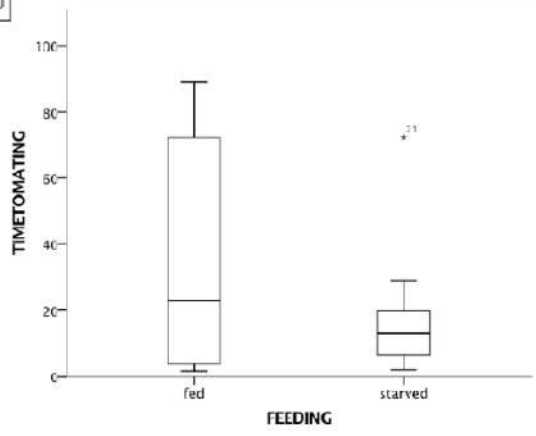
IPOTESI ALTERNATIVA: è diverso. È un TEST A DUE CODE.

VIOLAZIONE ASSUNZIONI – TEST NON PARAMETRICI – MW

PRESENTAZIONE MASCHIO A 24 FEMMINE
11 AFFAMATE E 13 NUTRITE
MONITORARE TEMPO ALL'ACCOPPIAMENTO

Feeding	Time	Feeding	Time
starved	1.9	fed	1.5
starved	2.1	fed	1.7
starved	3.8	fed	2.4
starved	9	fed	3.6
starved	9.6	fed	5.7
starved	13	fed	22.6
starved	14.7	fed	22.8
starved	17.9	fed	39
starved	21.7	fed	54.4
starved	29	fed	72.1
starved	72.3	fed	73.6
		fed	79.5
		fed	88.9

DISTRIBUZIONI NON NORMALI E NON 'NORMALIZZABILI'



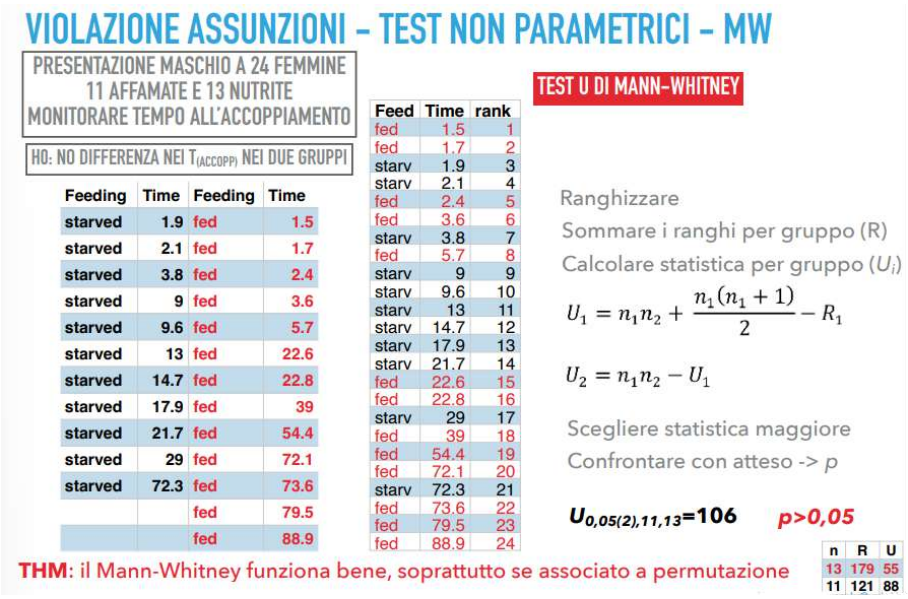
Come si vede dal box plot, le distribuzioni sono molto asimmetriche, con un outlier nelle femmine a digiuno. **Nessuno dei due gruppi ha una distribuzione simmetrica né è possibile fare una trasformazione (tipo logaritmica) che le porti ad una distribuzione normale**. Inoltre il campione è piccolo, quindi NON si può usare il teorema del limite centrale.

Con queste ipotesi, **NON POSSIAMO USARE IL TEST T A DUE CAMPIONI**.

Usiamo quindi il test U.

- Primo step: ordinare tutti i valori in ordine crescente, COMBINANDO I DATI DEI DUE GRUPPI (rosso e nero)
- Il **valore più piccolo** proviene dal gruppo rosso ed è 1,5, e avrà **RANGO 1**, il valore più grande viene sempre dal gruppo rosso ed è 88,9, e avrà RANGO 24.
- Ora calcoliamo la **SOMMA DEI RANGHI** per uno dei due gruppi: es. nero 3+4+7+9+10+11+12+13+14+17+21=121

- Ora usiamo le due somme per calcolare la nuova **grandezza, U, che è IL NUMERO DI VOLTE CHE UN VALORE DEL CAMPIONE NERO è MINORE DI UN VALORE DEL CAMPIONE ROSSO se confrontassimo tutte le possibili coppie**.



Feeding	Time	Feeding	Time
starved	1.9	fed	1.5
starved	2.1	fed	1.7
starved	3.8	fed	2.4
starved	9	fed	3.6
starved	9.6	fed	5.7
starved	13	fed	22.6
starved	14.7	fed	22.8
starved	17.9	fed	39
starved	21.7	fed	54.4
starved	29	fed	72.1
starved	72.3	fed	73.6
		fed	79.5
		fed	88.9

Feed	Time	rank
fed	1.5	1
fed	1.7	2
starv	1.9	3
starv	2.1	4
fed	2.4	5
fed	3.6	6
starv	3.8	7
fed	5.7	8
starv	9	9
starv	9.6	10
starv	13	11
starv	14.7	12
starv	17.9	13
starv	21.7	14
fed	22.6	15
fed	22.8	16
starv	29	17
fed	39	18
fed	54.4	19
fed	72.1	20
starv	72.3	21
fed	73.6	22
fed	79.5	23
fed	88.9	24

n	R	U
13	179	55
11	121	88

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

U1=88

- Calcolo poi anche U del gruppo rosso.

$$U_2 = n_1 n_2 - U_1$$

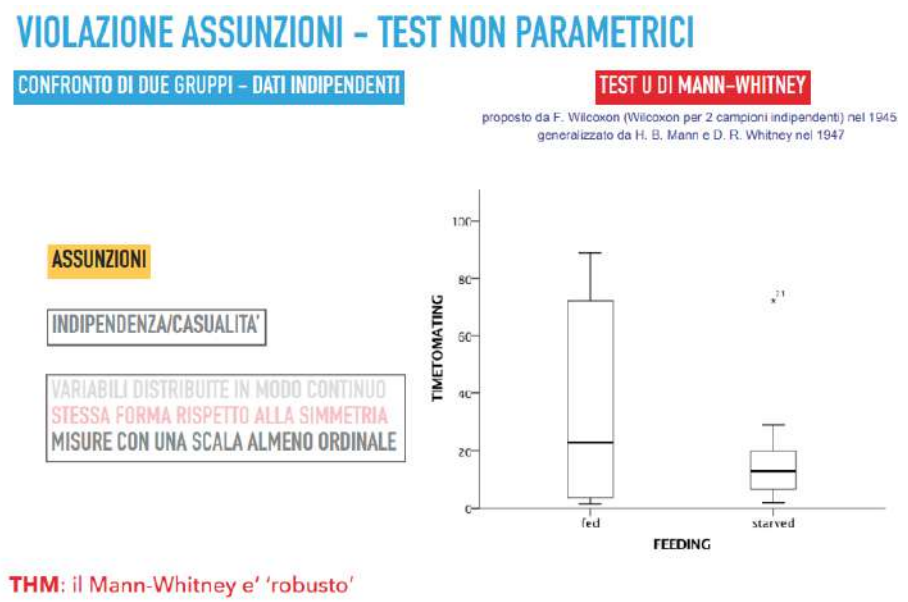
U2=55

- **SCEGLIAMO IL MAGGIORE TRA I DUE, che è U1=88.**
- Determiniamo infine il P value confrontando il valore di U calcolato con il valore critico della distribuzione nulla per U, che sono forniti dalle TAVOLE STATISTICHE: con **alpha=0,05 e n1=11 e n2=13 ho che U0,05=106**
- Il valore maggiore di U nostro è MINORE del valore di U delle tavole. Quindi NON possiamo RIFIUTARE L'IPOTESI NULLA.

Ma cosa succede se ci sono due valori uguali? Come assegno il rango? ASSEGNO IL RANGO MEDIO! Per esempio, se il due valori dovrebbero essere nella posizione 4 e 5, AD ENTRAMBI DO RANGO 4,5

TENIAMO BEN CONTO DEL FATTO CHE QUESTO TEST SE IL CAMPIONE è MOLTO GRANDE NON VIENE USATO PERCHE' POSSO APPROSSIMARE LA DISTRIBUZIONE CON UNA BUONA APPROSSIMAZIONE ALLA NORMALE STANDARD.

Assunzioni dei test non parametrici: NON si basano sulla distribuzione normale, ma comunque formulano alcune assunzioni, anche se meno. per esempio, ENTRAMBI I CAMPIONI DEVONO ESSERE ESTRATTI DALLE RISPETTIVE POPOLAZIONI. Non possono ovviamente essere casuali.



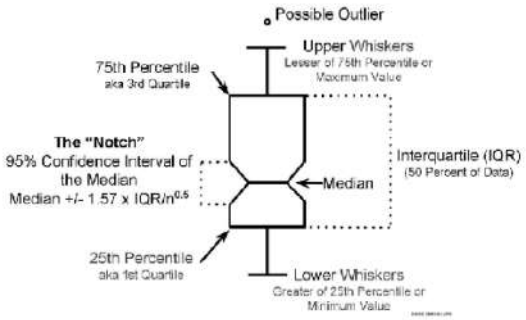
Il test dei ranghi di Wilcoxon, inoltre, assume che la distribuzione delle misure nella popolazione sia SIMMETRICA.

Il test U invece confronta le distribuzioni di due gruppi. VIENE USATO PER

CONFRONTARE LE TENDENZE CENTRALI DI DUE GRUPPI, ma serve l'assunzione che **LE DISTRIBUZIONI NEI DUE GRUPPI ABBIANO LA STESSA FORMA**: per esempio, STESSA VARIANZA E STESSO GRADO DI ASIMMETRIA.
I TEST NON PARAMETRICI SONO MENO POTENTI DI QUELLI PARAMETRICI

INTERVALLI DI CONFIDENZA NON PARAMETRICI

NOTCHED BOX-PLOT



Use `geom_boxplot()` and set `notch = TRUE`:

```
ggplot(dataset, aes(x = factor(factorname), y = variablename)) +
  geom_boxplot(notch = TRUE)
```

ESERCIZI

Compito di settembre

All'esame si vede una schermata con l'appello di riferimento
Due fasi

- Fase 1: NON VALUTATA, ma ci stimola ad applicare alla fine del corso le nostre competenze e conoscenze l'analisi di un articolo. Approccio critico standardizzato. La fase 1 si fa prima del primo tentativo che si dà, anche se si boccia l'esame.
- Fase 2: PRE SCREENING: visto che non vuole fare troppi orali, e dato che la natura del corso è pratica, nella prima fase dell'esame abbiamo delle domande a risposta multipla, che sono 15 domande di tipo teorico che fanno sì che si possano verificare le conoscenze teoriche (aka le definizioni). Per superare questa fase si deve superare il 50%. Ogni domanda a risposta multipla tranne quella di probabilità (vale 3 punti) vale 2 punti. Ci sono tra queste domande anche quelle di R.

  Esempio di domanda: considera quella falsa. Esempio per cosa vuol dire variabili indipendenti

  Se ho due risposte giuste il totale del punteggio viene diviso per due. Le risposte sbagliate paghi pegno. Siccome sono due giuste, si prende un punto e non due. A volte, infatti, si possono avere anche due risposte giuste e le devo METTERE ENTRAMBE

LEZIONE 17

Disegno sperimentale

Come stimare un n campionario sensato

Come funziona un disegno sperimentale?

INTERPRETAZIONE DELLE DOMANDE BIOLOGICHE

DISEGNI SPERIMENTALI – ANOVA – POWER ANALYSIS

ANOVA: tecnica identica al test t che si basa sul rapporto tra varianze e non differenze tra medie. Si usa per studiare il confronto tra medie di più di due gruppi. Alternative non parametriche per analisi della varianza.

Il disegno sperimentale più noto è quello dell'esperimento controllato in laboratorio. La cosa che distingue un campione dall'altro è il **TRATTAMENTO**. Per esempio, con i topi esistono delle linee clonali identiche se non per una singola mutazione. MANIPOLAZIONE dello scenario se cambi le condizioni per vedere la differenza tra trattamenti e controlli

Disegno QUASI sperimentale: si ha comunque manipolazione ma le condizioni NON sono del tutto controllate.

Studi OSSERVAZIONALI: non c'è per nulla controllo: si raccolgono le misure delle variabili di interesse e si osserva per esempio se sono correlate. **Il problema è che da studi osservazionali in cui non ci sono condizioni controllate non si possono tenere sotto controllo le variabili esterne, che sono dette VARIABILI DI CONFONDIMENTO**. Di solito si combinano questi tipi di approcci per riuscire a mettere a punto un disegno sperimentale sensato

Una variabile di confondimento è una variabile che maschera o distorce la relazione causa-effetto tra le variabili di interesse in uno studio.

Perché si fa quindi un disegno di ricerca

La cosa principale che si fa per **ridurre la DISTORSIONE è la RANDOMIZZAZIONE**, mentre per ridurre l'errore di campionamento, che dipende da n campionario e quindi dall'errore di primo tipo (rapporto tra deviazione standard e radice di n), bisogna **RIDURRE LA DEVIAZIONE STANDARD O AUMENTARE N CAMPIONARIO**

Un disegno sperimentale prevede che ci siano dei soggetti che vengano attribuiti casualmente. Identifico almeno due gruppi e riduco il rischio di una distorsione che l'operatore non sa ridurre. Quando è possibile aumentare le repliche di un numero adeguato, ma a volte non si può fare perché mancano campioni biologici.

Esempio: riduzione dell'infezione da HIV

Lo standard di riferimento dei disegni sperimentali è infatti la **SPERIMENTAZIONE CLINICA**, per cui sempre ci sono dei soggetti trattati e dei controlli. Alla fine dell'esperimento, 59 delle 376 donne trattate con il principio attivo erano positive, rispetto alle 45 su 389 che invece non avevano il principio attivo. quindi l'odds di contrarre il virus era più alto nel gruppo trattato rispetto al controllo. I soggetti sperimentali NON devono sapere chi era il trattamento e chi il controllo, e neanche i ricercatori. **Questo è importantissimo per la ricerca, perché non sapere chi è trattato e chi è controllo evita di influenzare sui risultati ottenuti**.

Questo tipo di studio è QUASI sperimentale, perché non controlli altro che la professione e il trattamento, NON è sperimentale perché in questo tipo di studi **i soggetti NON POSSONO ESSERE ISOLATI**; quindi, ci sono altri fattori non strettamente legati al trattamento che influiscono sul risultato. Se fossero tutti chiusi in laboratorio sarebbe sperimentale

Cosa ci chiederebbe: rappresentare graficamente i dati, usare il test giusto e ci chiede se questo esperimento conferma o confuta l'ipotesi che il gel funzioni o no.

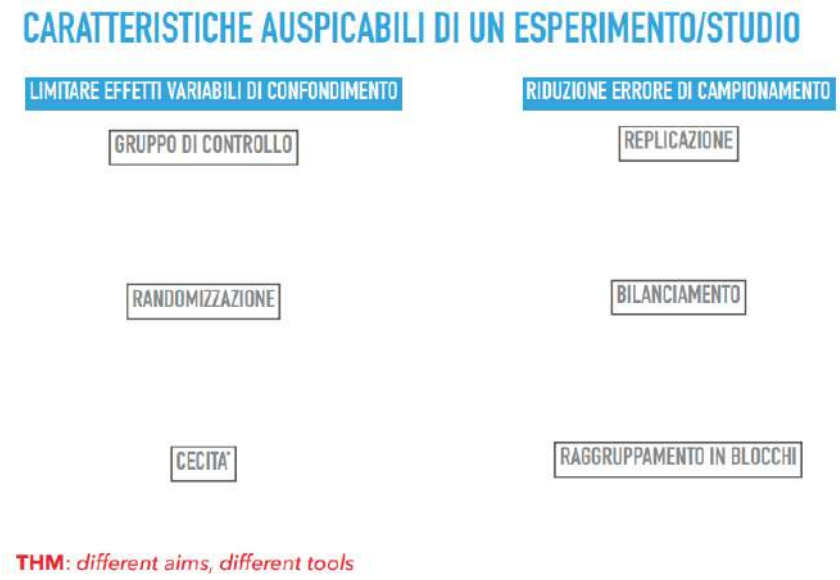
Rappresentazione: MOSAIC PLOT che viene pressoché identico: sulle ordinate la FREQUENZA RELATIVA, sulle ascisse i due casi.

Ho comunque delle stime, e quindi posso calcolare l'INTERVALLO DI CONFINDENZA, descrivendo i valori con un punto e il loro intervallo di confidenza.

Ipotesi nulla: non ci sono differenze.

Di quanto diminuisce il rischio per soggetti trattati e non trattati è una buona domanda che ci si può fare. Si deve calcolare **l'ODDS RATIO**

La randomizzazione serve per ridurre le variabili di confondimento. **Attribuisco poi dei valori random e assegno le variabili a diversi gruppi**.



Il gruppo di controllo deve dare il gel al pari del gruppo dei trattati, se no potrebbero esserci dei danni ulteriori al trattamento dovuti al fatto che prende il gel e potrebbe fare male

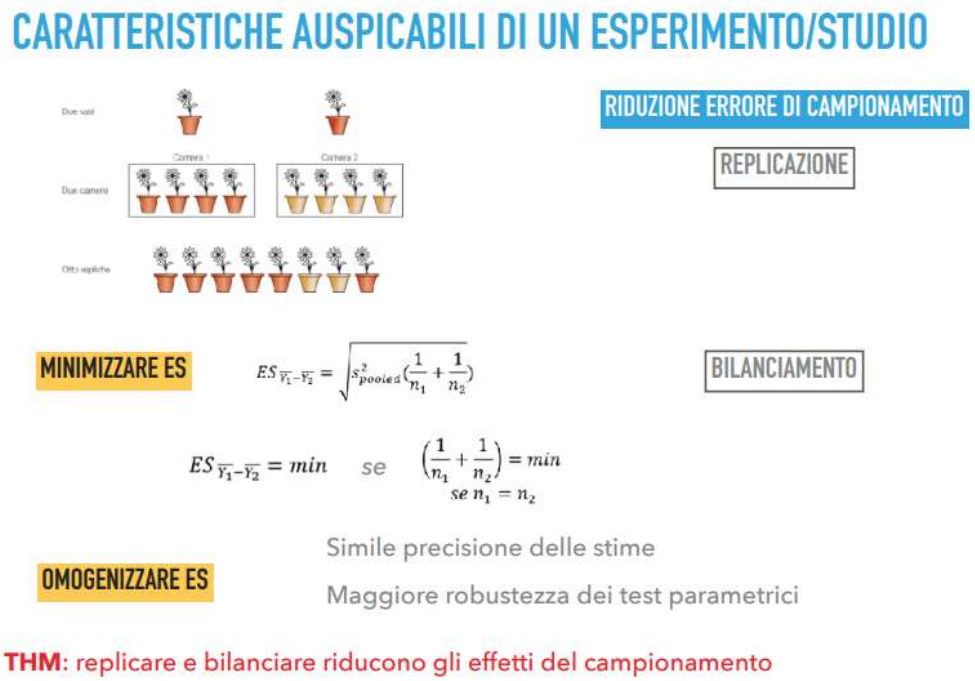
SCOPO DEL DISEGNO SPERIMENTALE: ridurre la distorsione nella stima e nella verifica degli effetti del trattamento e ridurre gli effetti dell'errore di campionamento

RIDURRE LA DISTORSIONE: considero un GRUPPO DI CONTROLLO, poi uso

RANDOMIZZAZIONE e ho la CECITA'. RIDURRE GLI EFFETTI DELL'ERRORE DI CAMPIONAMENTO: REPLICAZIONE, BILANCIAMENTO (stesso numero di individui per i due gruppi), RAGGRUPPAMENTO IN BLOCCHI (soggetti nell'esempio sopra raggruppati sulla base dell'ambulatorio che frequentavano). Questo ha senso se l'ambiente tra i blocchi è uniforme. Ogni blocco è una replica. È come si disegna per dati appaiati, l'n campionario non è per due, è la singola coppia.

RIDURRE GLI EFFETTI DELL'ERRORE DI CAMPIONAMENTO

Un modo per evitare gli errori di campionamento è RENDERE COSTANTI LE CONDIZIONI SPERIMENTALI.



Negli esperimenti sul campo potrebbe essere difficile obv.

La **REPLICAZIONE** è **l'applicazione di ogni trattamento per più unità sperimentali INDIPENDENTI**

Esempio nella slide: si vedono tre disegni sperimentali usati per confrontare la crescita di due piante sottoposte a due trattamenti di temperatura indicati dal colore dei vasi. Solo il disegno più in basso indice che c'è stata una replicazione: le unità sperimentali sono infatti state sparpagliate rispetto al trattamento che hanno ottenuto!

L'UNITA' SPERIMENTALE può essere o la singola pianta/singolo individuo oppure insieme di singoli individui che sono considerati un gruppo unico (capsula petri, gabbia di animali…)

È importante avere più replicazioni perché si RIDUCE L'ERRORE STANDARD, che dipende infatti da n al denominatore.

Un altro aspetto importante per ridurre l'errore di campionamento è Il **BILANCIAMENTO**, che si ha quando **TUTTI I TRATTAMENTI HANNO UGUALE DIMENSIONE CAMPIONARIA**. Sappiamo infatti che l'errore standard è minimo quando è **minima la somma** $\left(\frac{1}{n_1}+\frac{1}{n_2}\right)$ **, e questo avviene SE n1 e n2 SONO UGUALI.**

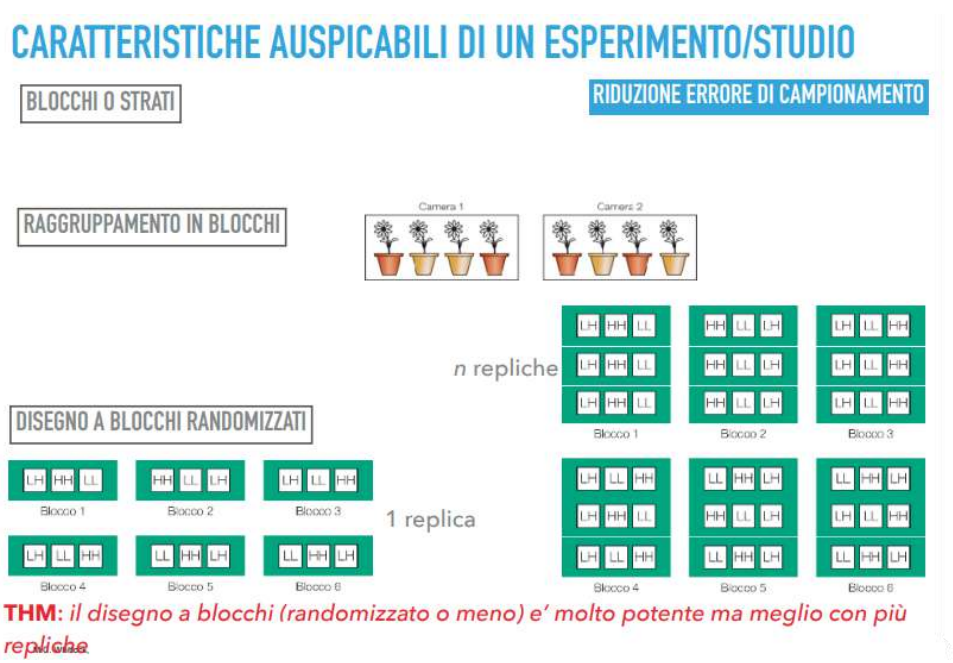
Altro aspetto importante, sempre per ridurre l'errore di campionamento, è il **RAGGRUPPAMENTO IN BLOCCHI**, che è una strategia molto utile per tenere conto della variabilità non dovuta al trattato. Le unità sperimentali vengono quindi divise in **BLOCCHI o STRATI** (si parla infatti di CAMPIONAMENTI STRATIFICATI), e in ogni blocco le singole unità condividono le stesse condizioni generali dell'ambiente sperimentale. Nei blocchi si attribuisce in modo random trattato e controllo: **OGNI BLOCCO HA QUINDI SIA TRATTATI CHE CONTROLLI**. (vedi vasi di fiori, due gruppi ciascuno con tutti e due i trattamenti).

IN QUESTO MODO SI ELIMINA LA VARIABILITA' CHE DERIVA DALLA DIFFERENZA TRA I BLOCCHI!

Un esempio di raggruppamento in blocchi è il disegno per DATI APPAIATI per due trattamenti: DUE TRATTAMENTI SONO APPLICATI A CIASCUN PLOT, CHE RAPPRESENTA UN BLOCCO!

Il disegno a BLOCCHI RANDOMIZZATI è MOLTO SIMILE AL DISEGNO PER DATI APPAIATI MA HA PIU' DI DUE TRATTAMENTI

Esempio: cavità degli alberi per esaminare se la quantità di lettiera di foglie in decomposizione al loro interno se influenzasse il numero di uova di insetti e la loro sopravvivenza. La lettiera è l'unica fonte di nutrienti. In un trattamento (LL) viene fornita una bassa quantità di lettiera, poi (HH) altissima quantità e poi (LH) inizialmente bassa ma poi viene aumentata dopo la deposizione delle uova. Viene usato un disegno a blocchi randomizzati, in cui le cavità arboree erano disposte in terne (i BLOCCHI), e ogni terna aveva un LL, un HH e un LH. La posizione di ciascun trattamento all'interno del blocco è stata randomizzata!

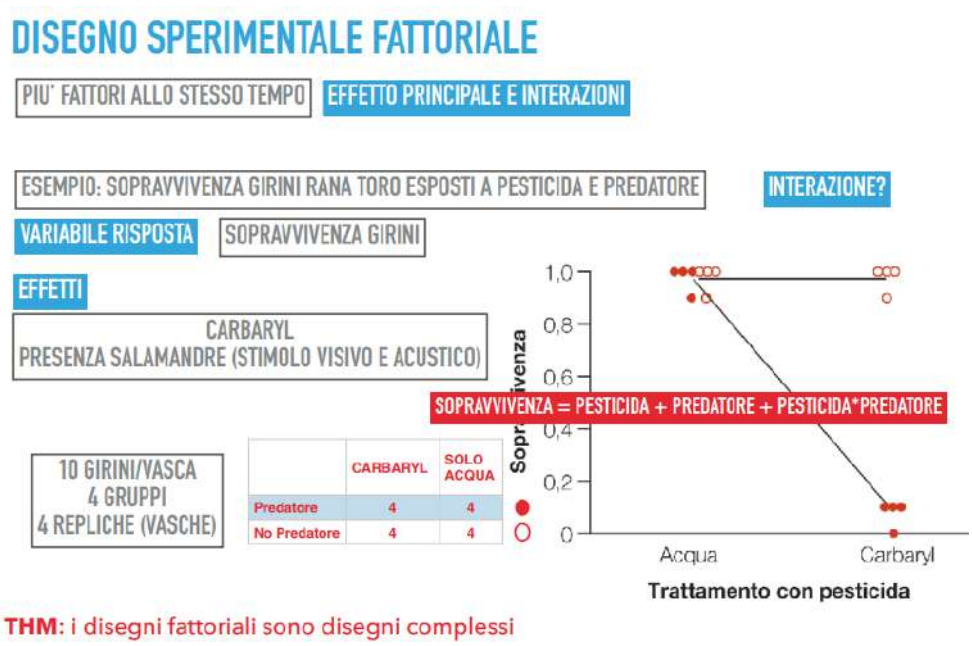


Questo è un **DISEGNO A BLOCCHI RANDOMIZZATI**: come un disegno a dati appaiati, **gli effetti del trattamento sono misurati dalle differenze tra i trattamenti solo DENTRO i blocchi. In questo modo si RIDUCE L'INFLUENZA DELLA VARIABILITA' TRA I BLOCCHI**.

ESPERIMENTI CON PIU' DI UN FATTORE

Un **FATTORE** è una variabile trattamento i cui effetti interessano il ricercatore. Spesso in uno stesso esperimento si usano più fattori perché è molto più vantaggioso poter rispondere a più domande con un singolo esperimento.

Inoltre, i fattori possono INTERAGIRE. Per esempio, le attività antropiche hanno causato un aumento di CO_2 e della temperatura atmosferica, e anche un incremento della deposizione di azoto. è stato possibile verificare che l'aumento di ognuno di questi fattori separatamente stimola la crescita delle piante, ma se questi fattori vengono considerati insieme i loro effetti congiunti possono essere molto diversi. **Nel DISEGNO FATTORIALE vengono studiate tutte le possibili combinazioni di trattamenti con due o più fattori**.



Esempio slide: pesticida che influenza la sopravvivenza dei girini. L'effetto del pesticida dipende dalla presenza di un predatore autoctono? Viene isolato

il predatore in gabbia in mood che non poteva causare un effetto diretto ma emetteva stimoli visivi e chimici. L'esperimento è stato condotto in vasche ciascuna con 10 girini. **Le quattro possibili combinazioni di due trattamenti con predatori (una vasca con i predatori e una senza) e con pesticida (pesticida o solo acqua**) sono state assegnate in modo casuale alle vasche, e per ogni combinazione erano previste **REPLICHE**. Si vede nel grafico della slide l'INTERAZIONE TRA GLI EFFETTI DEI TRATTAMENTI CON PESTICIDA E CON PREDATORE: ogni puntino rappresenta la frazione di girini sopravvissuti nella vasca

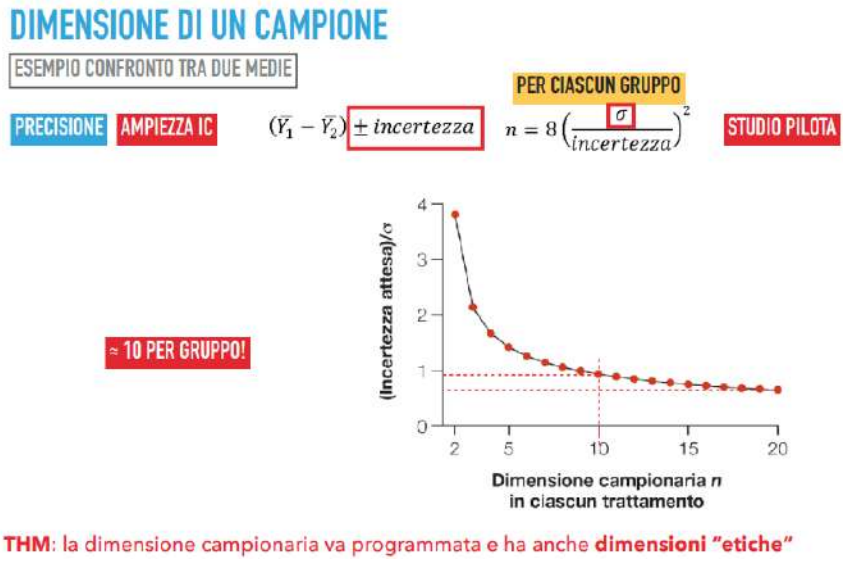
IN QUESTO CASO L'UNITA' SPERIMENTALE è LA VASCA, NON IL SINGOLO GIRINO. La sopravvivenza si vede essere alta tranne quando il pesticida era usato con il predatore. Quindi l'effetto di una variabile era influente sull'altra. UN ESPERIMENTO CHE TRATTAVA SOLO IL PREDATORE O SOLO IL PESTICIDA NON AVREBBE PERMESSO DI OSSERVARE NESSUN EFFETTO.

Cosa fare quando NON SI POSSONO CONDURRE ESPERIMENTI?

Si fanno gli STUDI OSSERVAZIONALI. L'idea è usare strategie di osservazione che riducano al minimo la DISTORSIONE (tipo con gruppi di controllo e cecità) e anche l'impatto dell'errore di campionamento (replicazione e bilanciamento). **NON SI PUO' FARE RANDOMIZZAZIONE**!

Due strategie negli esperimenti osservazionali è l'APPAIAMENTO: ogni individuo nel gruppo di trattamento viene appaiato con un individuo di controllo con valori identici o molto simili in relazione alle possibili variabili che potrebbero far confondere. Es. **ogni individuo con una certa malattia viene appaiato con uno sano per tutte le caratteristiche che potrebbero influenzare lo studio, come peso, altezza, età, sesso e background etnico**.

La seconda strategia è l'**AGGIUSTAMENTO**, per cui si usano metodi statistici come l'analisi della **COVARIANZA** per correggere la differenza tra gruppi di trattamento e gruppi di controllo.



Un altro aspetto importante nella pianificazione di un esperimento è la **SCLETA DELLA DIMENSIONE DEL CAMPIONE**

Non ha senso condurre uno studio se il campione ha dimensioni troppo ridotte, o se l'intervallo di confidenza è troppo ampio a causa di n piccolo. può poi essere vantaggioso anche studiare troppi individui, perché ogni replica è tempo e denaro o anche se lo studio non è sicuro per i gruppi coinvolti.

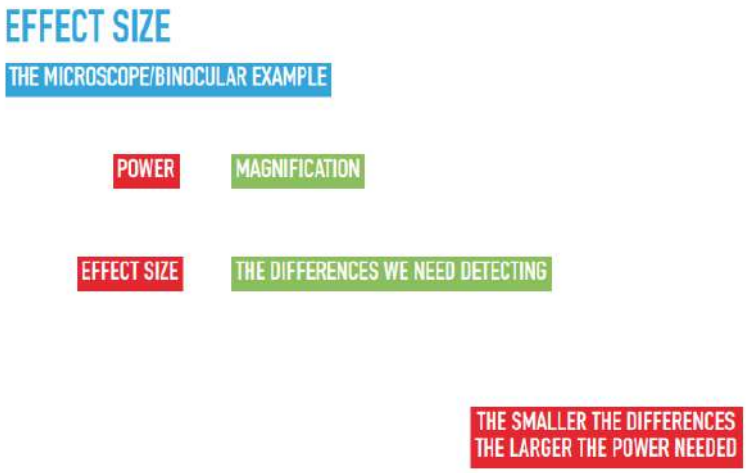
Come si sceglie? Se si vuole

PRECISIONE si agisce in un modo, se si vuole POTENZA in un altro.

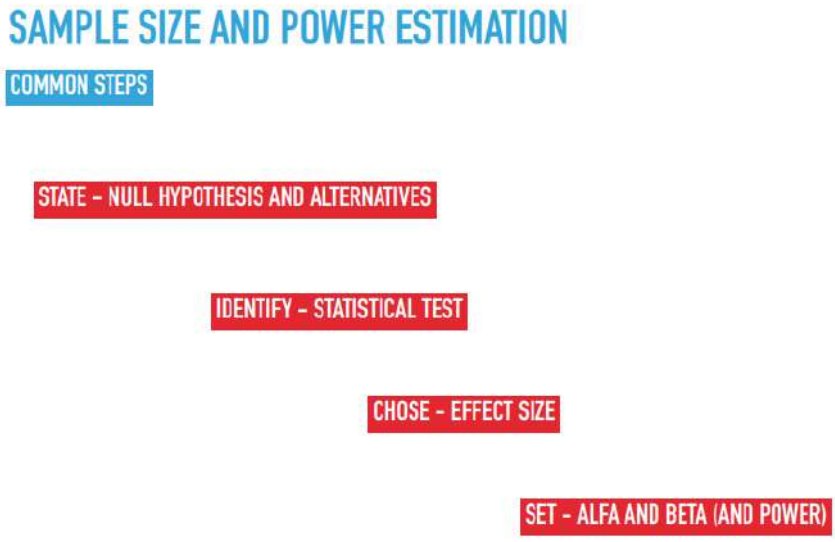
Pianificare la PRECISIONE

Bisogna scegliere una dimensione di un campione che fornisca un intervallo di confidenza non troppo ampio, ma compatibilmente con lo sforzo sperimentale. Si vede nel grafico della slide la precisione attesa dell'intervallo di confidenza al 95% della differenza tra le medie di due trattamenti in funzione della dimensione campionaria del trattamento. Si vede bene come dimensioni campionarie ridotte portino ad intervalli stimati della differenza tra le medie nei trattamenti molto ampi. La incertezza a n=20 scende moltissimo per n=200 (10 per gruppo)!

FORMULE IMPORTANTISSIME NELLA SLIDE PER CALCOLARE L'n PERFETTO! INCERTEZZA è META' DELL'AMPIEZZA DELL'INTERVALLO DI CONFIDENZA. BISOGNA MENTRE SI PIANIFICA SCEGLIERE L'INCERTEZZA (QUALE DIFFERENZA TRA LE DUE MEDIE VOGLIO RILEVARE) **CHE SI PUO' ACCETTARE IN ANTICIPO. DA LI' QUINDI SI PUO' CAPIRE QUANTI N SERVONO CON LA FORMULA SOPRA.**



La **differenza che voglio essere in grado di rilevare è spesso indicata come EFFECT SIZE:** se si vuole lavorare su umani e primati e non invertebrati o tessuti, è importante fare analisi approfondite in fase pre-sperimentale.



Esiste tutta una serie di software, tra cui **G power**, che permette di dirti che test stai facendo, che tipo di test vuoi fare e che tipo di analisi. Per esempio, voglio fare un'analisi a priori o a posteriori, cioè prima o dopo l'esperimento. Poi dici che vuoi il test a due code, e quindi devo AUMENTARE N CAMPIONARIO se voglio che diminuisca l'errore standard. questo software calcola l'n campionario minimo che serve. Con questo giochino possiamo capire quali sono i parametri che determinano l'n campionario.

DIMENSIONE DI UN CAMPIONE

ESEMPIO CONFRONTO TRA DUE MEDIE

PROBABILITA' DI RIFIUTARE H_0 QUANDO E' FALSA

$potenza = 1 - \beta$; *dove* $\beta = errore\ II\ tipo$ $\quad \beta = 0{,}20\ e \rightarrow potenza = 0{,}80$

POTENZA

σ $\quad (\overline{Y_1} - \overline{Y_2})$ $\quad |\mu_1 - \mu_2| \geq D$ $\quad n \approx 16\left(\frac{\sigma}{D}\right)^2$

Software: G-Power (http://www.gpower.hhu.de/en.html)

Online tools Power and Sample Size .com
http://powerandsamplesize.com/Calculators/

PERDITA DI DATI Prudenza! Meglio (*n + qualcosa*)

THM: il calcolo della potenza del test dato *n* e' difficile. Varia tra test e disegni sperimentali

Se non si ha un risultato significativo non si può dire che ci sono differenze, ma anche che non ci sono differenze, perché devo avere la **POTENZA ALMENO DI 0,8**, e se non lo è non posso trarre conclusioni affrettate.

Un 10/15% di campioni biologici si perde per vari motivi; quindi, è bene essere prudenti e aumentare il nostro campione auspicabile attraverso l'analisi della potenza e la stima delle misure del campione. **È legittimo dire che la propria analisi della potenza indica un numero di soggetti n, ma in previsione di una eventuale perdita di campioni stimata a m chiedo il doppio dei campioni.** Obv non sempre succede, se chiedo cervelli umani è ovvio che non ne chiedo di più.

Quindi questo software mi permette una STIMA DI N E DELLE POTENZA!

T TEST

1. State the null hypothesis and whether the alternative hypothesis is one- or two-sided.
2. Estimate the effect size (E) as the difference in the mean value of the continuous variable between the study groups.
3. Estimate variability as the standard deviation (S) of that variable.
4. Calculate the standardized effect size (E/S), defined as the effect size divided by the standard deviation of the outcome variable.
5. Set α and β.

TABLE 6A SAMPLE SIZE *PER GROUP* FOR COMPARING TWO MEANS

ONE-SIDED $\alpha =$		0.005			0.025			0.05	
TWO-SIDED $\alpha =$		0.01			0.05			0.10	
E/S^*	$\beta =$ 0.05	0.10	0.20	0.05	0.10	0.20	0.05	0.10	0.20
0.10	3,565	2,978	2,338	2,600	2,103	1,571	2,166	1,714	1,238
0.15	1,586	1,325	1,040	1,157	935	699	963	762	551
0.20	893	746	586	651	527	394	542	429	310
0.25	572	478	376	417	338	253	347	275	199
0.30	398	333	262	290	235	176	242	191	139
0.40	225	188	148	164	133	100	136	108	78
0.50	145	121	96	105	86	64	88	70	51
0.60	101	85	67	74	60	45	61	49	36
0.70	75	63	50	55	44	34	45	36	26
0.80	58	49	39	42	34	26	35	28	21
0.90	46	39	32	34	27	21	28	22	16
1.00	38	32	26	27	23	17	23	18	14

*ES is the standardized effect size, computed as E (expected effect size) divided by S (SD of the outcome variable). To estimate the sample size, read across from the *standardized effect* size, and down from the specified values of α and β for the required sample size in each group. For a one-sample *t* test, the total sample size is one-half of the number listed.

Esempio: procedura di calcolo della potenza di un t test per confrontare la media di due gruppi di campioni indipendenti. Chiarire a priori nel nostro progetto se, secondo noi, l'ipotesi alternativa è a una coda o due. Poi la cosa che faccio è immaginare di avere una variabile di risposta continua, e poi fare l'ipotesi o a partire

EXERCISE – T TEST

EXAMPLE 6.1 Sample Size When Using the *t* Test

Problem: The research question is whether there is a difference in the efficacy of albuterol and ipratropium bromide for the treatment of asthma. The investigator plans a randomized trial of the effect of these drugs on FEV_1 (forced expiratory volume in 1 second) after 2 weeks of treatment. A previous study has reported that the mean FEV_1 in persons with treated asthma was 2.0 liters, with a standard deviation of 1.0 liter. The investigator would like to be able to detect a difference of 10% or more in mean FEV_1 between the two treatment groups. How many patients are required in each group (albuterol and ipratropium) at α (two-sided) = 0.05 and power = 0.80?

The ingredients for the sample size calculation are as follows:

1. Null Hypothesis: Mean FEV_1 after 2 weeks of treatment is the same in asthmatic patients treated with albuterol as in those treated with ipratropium.
   Alternative Hypothesis (two-sided): Mean FEV_1 after 2 weeks of treatment is different in asthmatic patients treated with albuterol from what it is in those treated with ipratropium.
2. Effect size = 0.2 liters (10% × 2.0 liters).
3. Standard deviation of FEV_1 = 1.0 liter.
4. Standardized effect size = effect size ÷ standard deviation = 0.2 liters ÷ 1.0 liter = 0.2.
5. α (two-sided) = 0.05; $\beta = 1 - 0.80 = 0.20$. (Recall that $\beta = 1 -$ power.)

da dati pregressi o facendo una ipotesi sul valore della frequenza. Per esempio, se so dalla letteratura che la media di un processo è intorno ai 3,9 voglio capire se c'è un processo che la discosta di 0,5 o di più. Questo vuol dire che un cambiamento di 0,4 non viene rilevato, perché io lo voglio minimo del 0,5. Questo è il punto del p VALUE perché, se non c'è significatività non dico che non ci sono differenze, ma che ci sono differenze inferiori a quelle che ho programmato.

TABLE 6A SAMPLE SIZE *PER GROUP* FOR COMPARING TWO MEANS

ONE-SIDED α =		0.005			0.025			0.05	
TWO-SIDED α =		0.01			0.05			0.10	
*E/S**	β = 0.05	0.10	0.20	0.05	0.10	0.20	0.05	0.10	0.20
0.10	3,565	2,978	2,338	2,600	2,103	1,571	2,166	1,714	1,238
0.15	1,586	1,325	1,040	1,157	935	699	963	762	551
0.20	893	746	586	651	527	394	542	429	310
0.25	572	478	376	417	338	253	347	275	199
0.30	398	333	262	290	235	176	242	191	139
0.40	225	188	148	164	133	100	136	108	78
0.50	145	121	96	105	86	64	88	70	51
0.60	101	85	67	74	60	45	61	49	36
0.70	75	63	50	55	44	34	45	36	26
0.80	58	49	39	42	34	26	35	28	21
0.90	46	39	32	34	27	21	28	22	16
1.00	38	32	26	27	23	17	23	18	14

**E/S* is the standardized effect size, computed as *E* (expected effect size) divided by *S* (SD of the outcome variable). To estimate the sample size, read across from the *standardized effect size*, and down from the specified values of α and β for the required sample size in each group. For a one-sample *t* test, the total sample size is one-half of the number listed.

Voglio vedere il flusso respiratorio dopo due settimane di trattamento. **È stato osservato che il FEV medio si ha da studi precedenti, se no posso fare uno studio pilota io.** comunque in questo caso **so sia la media che la deviazione standard**. Ora mi chiedo il **numero di pazienti per l'esperimento perché, se non lo so non posso andare avanti**. non si sa quale dei due farmaci funzioni meglio: ESPERIMENTO A DUE CODE!

Flusso medio 2L

Deviazione standard 1L

Vogliono rilevare il 10% di differenza nel flusso respiratorio forzato di almeno il 10%. Quanti campioni servono è la domanda.

Due code, 0,05 alfa, 0,2 beta

Bisogna **QUANTIFICARE L'EFFECT SIZE**, misurare la deviazione standard, stimare l'effect size standardizzato.

Effect size: 0,2, perché è il 10% DEL FLUSSO MEDIO

Effect size standardizzato: divido l'effect size per la deviazione standard, quindi di nuovo 0,2

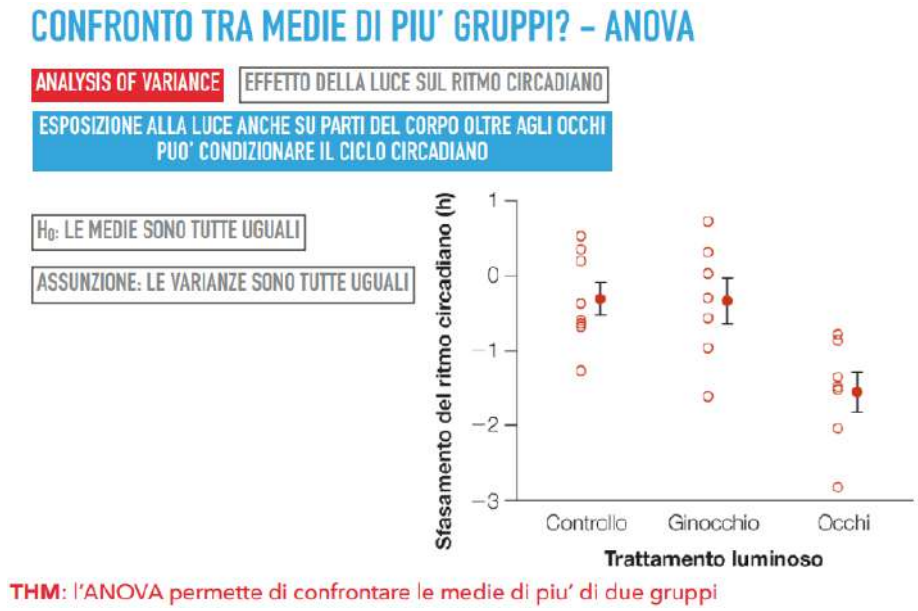
Ipotesi nulla è che la media per i due farmaci sia uguale, quella alternativa è che le due medie siano diverse nei due gruppi.

Ci servono 394 pazienti per gruppo.

Supponi che non hai i soldi per quel numero così elevato, NON puoi cambiare la media di riferimento, ma forse posso cambiare la deviazione standard, cioè, vedere se esistono dei campioni meno variabili: riduzione della variabilità del campione iniziale. Facendo così, ho un sottogruppo di pazienti, e quindi devo rivalutare la media. però in assenza di questa cosa posso variare l'effect size! Per esempio, lo faccio al 20% e non al 10%, così vedo differenze SOLO se sono TANTO significative.

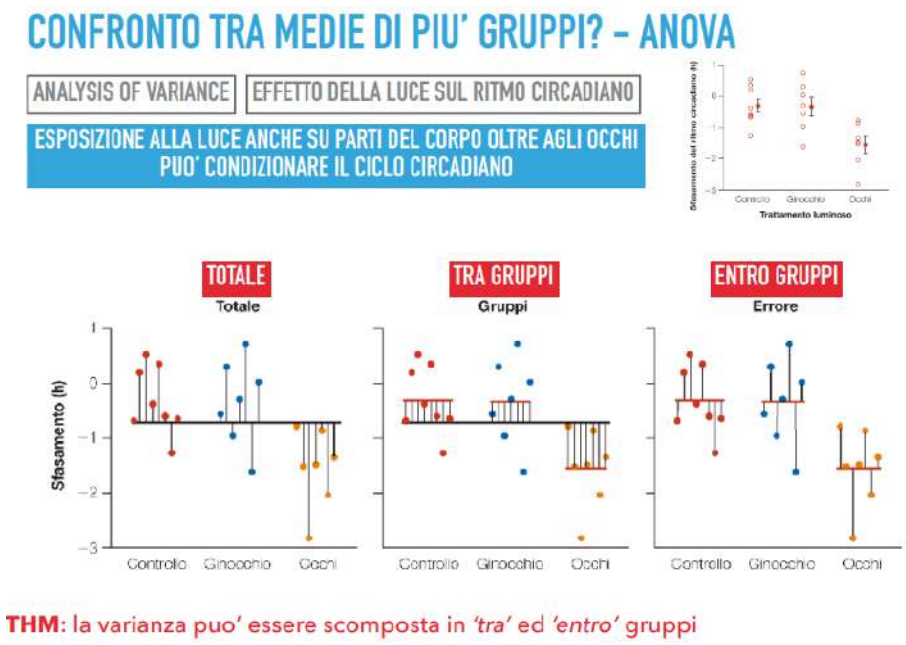
La cosa che vuoi è verificare che i farmaci abbiano il loro effetto. se decido di fare un cambiamento del 15% ho 0,3.

Quando ho TRE GRUPPI e non ho un computer per calcolare il p value, o faccio dei TEST T MULTIPLI CORREGGENDO ALFA CON BONFERRONI. Per farlo in maniera più efficiente si fa l'**ANOVA**
Posso usare l'ANOVA anche per un confronto tra due gruppi semplice, aka con il t test.
L'ANOVA è l'ANALISI DELLA VARIANZA, e si usa per valutare simultaneamente se le MEDIE DI K GRUPPI SONO UGUALI O NO. permette quindi di **verificare se INDIVIDUI APPARTENENTI A DIVERSI GRUPPI SIANO IN MEDIA PIU' DIVERSI O NO RISPETTO A QUELLI APPARTENENTI ALLO STESSO GRUPPO**.



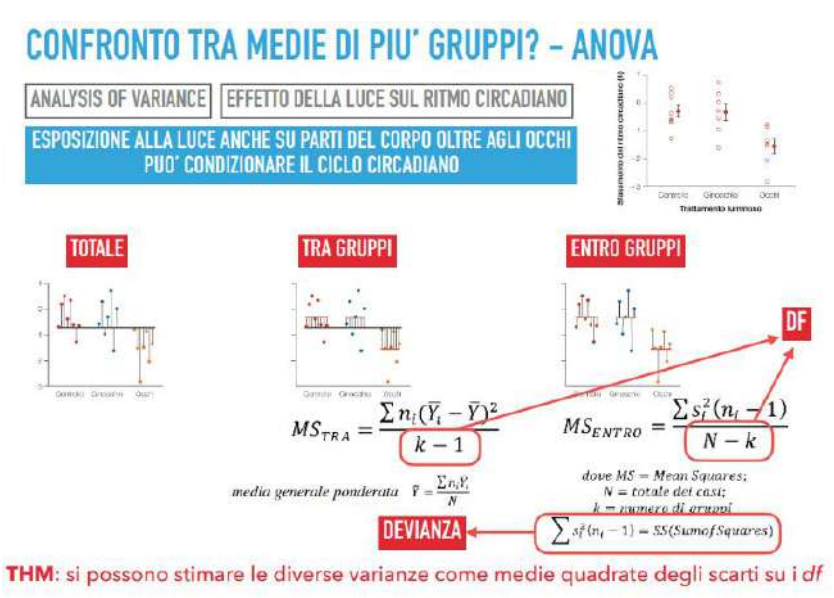
Esempio: jet lag: si supera quando l'alternarsi di luce e buio si sincronizza di nuovo con il ritmo circadiano interno che è stato sfasato. Esperimento in cui viene misurato il ritmo circadiano valutando la produzione di melatonina in 22 soggetti con tre diversi trattamenti luminosi. Sono stati svegliati dal sonno e soggetti a 3 ore di illuminazione intensa o solo sugli occhi, o solo sulle ginocchia o in nessuno dei due (controllo). Gli effetti del trattamento sono stati visti il giorno dopo: misura positiva dello sfasamento vuol dire anticipo, negativa vuol dire ritardo.
IPOTESI NULLA TEST ANOVA: **medie delle popolazioni sono uguali in tutti i trattamenti. Se la popolazione i esima ha come media μ_i allora ho che:**
IPOTESI NULLA : $\mu_i = \mu_j$ per ogni i, j
IPOTESI ALTERNATIVA: almeno una μ_i è diversa: **ALMENO UNA, NON TUTTE DIVERSE**!

ASSUNZIONE: le varianze sono tutte uguali.



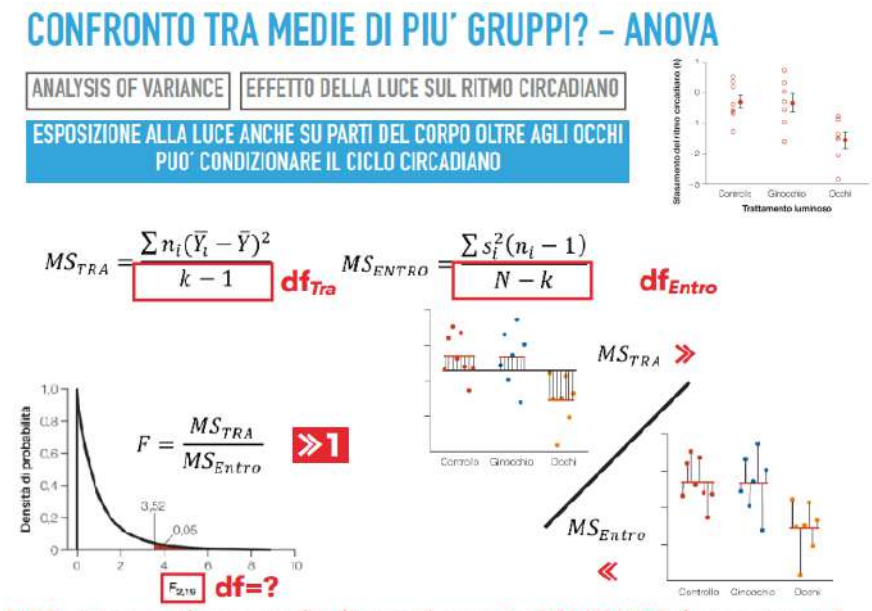
Anche nel caso in cui le medie di tutte le popolazioni fossero tutte uguali, comunque le MEDIE CAMPIONARIE sarebbero diverse. Quindi il **ruolo dell'ANOVA è quello di determinare se la variazione tra le medie campionarie dei gruppi sia MAGGIORE di quella ATTESA PER IL CASO**.
L'ANOVA confronta due fonti di variabilità nei dati. Esempio nella slide: i dati sono stati dispersi orizzontalmente in modo che si possano vedere (non sull'asse verticale, dove obv sono rimasti uguali). il segmento verticale rappresenta lo scarto totale di ogni dato dalla media generale DI

TUTTI I DATI. Se infatti elevassimo gli scarti al quadrato e ne calcolassimo la media, otterremmo la varianza campionaria di tutti i dati insieme. Con l'ANOVA andiamo invece a dividere lo scarto totale di ogni osservazione in due parti, che sono le fonti di variabilità dei dati: **UNO è LO SCARTO DI OGNI OSSERVAZIONE DALLA MEDIA DEL SUO GRUPPO (RESIDUO o ERRORE), L'ALTRO è LO SCARTO DELLA MEDIA DEL GRUPPO A CUI L'OSSERVAZIONE APPARTIENE RISPETTO ALLA MEDIA GENERALE**. Se eleviamo al quadrato gli "errori" o "residui" e calcoliamo la **media (ponderata)** dei risultati otteniamo la varianza campionaria comune: parliamo di **VARIANZA ENTRO I GRUPPI (o MEDIA DEI QUADRATI DEGLI ERRORI)**. Invece la **MEDIA DEI QUADRATI TRA I GRUPPI (o VARIANZA TRA I GRUPPI)** rappresenta **la variabilità tra individui appartenenti a diversi gruppi**.



Se l'ipotesi nulla fosse vera e le medie delle popolazioni nei gruppi fossero tutte uguali, allora le due medie dei quadrati dovrebbero essere circa uguali. così la variabilità entro i gruppi non sarebbe diversa da quella tra i gruppi, se non per il CASO. Se invece la media dei quadrati TRA i gruppi è significativamente maggiore della media dei quadrati degli errori, allora vuol dire che esistono differenze reali tra le medie delle popolazioni.

Quindi l'ANOVA è UN TEST CHE CONFRONTA QUESTE DUE MEDIE DI QUADRATI (MStra e MSentro)



MEDIA DEI QUADRATI DEGLI ERRORI: mean square dell'errore: MSerrore o MSentro

$$MS_{errore} = \frac{\sum s_i^2(n_i - 1)}{N - k}$$

Gli s_i nella formula sono le DEVIAZIONI STANDARD CAMPIONARIE OTTENUTE DA CIASCUN GRUPPO, mentre n_i sono le DIMENSIONI DEL CAMPIONE, mentre $n_i - 1$ è il numero di gradi di libertà dell'i-esimo gruppo. Il denominatore è la domma dei gradi di libertà per i differenti gruppi. N è il numero totale delle osservazioni nell'insieme dei gruppi

MEDIA DEI QUADRATI TRA GRUPPI: mean square tra gruppi: MSgruppi

$$MS_{gruppi} = \frac{\sum n_i(\bar{Y}_i - \bar{Y})^2}{k-1}$$

Con k numero TOTALE dei gruppi. Il denominatore di questa formula, k-1, p il numero di gradi di libertà per i gruppi.

Infine, a **MEDIA GENERALE $\bar{Y}$** è la media di tutti i dati di tutti i gruppi. Non è uguale alla media delle medie campionarie di ciascun gruppo, perché può succedere che **LE MEDIE CAMPIONARIE SIANO DIVERSE L'UNA DALL'ALTRA**.

$$\bar{Y} = \frac{\sum n_i \bar{Y}_i}{N}$$

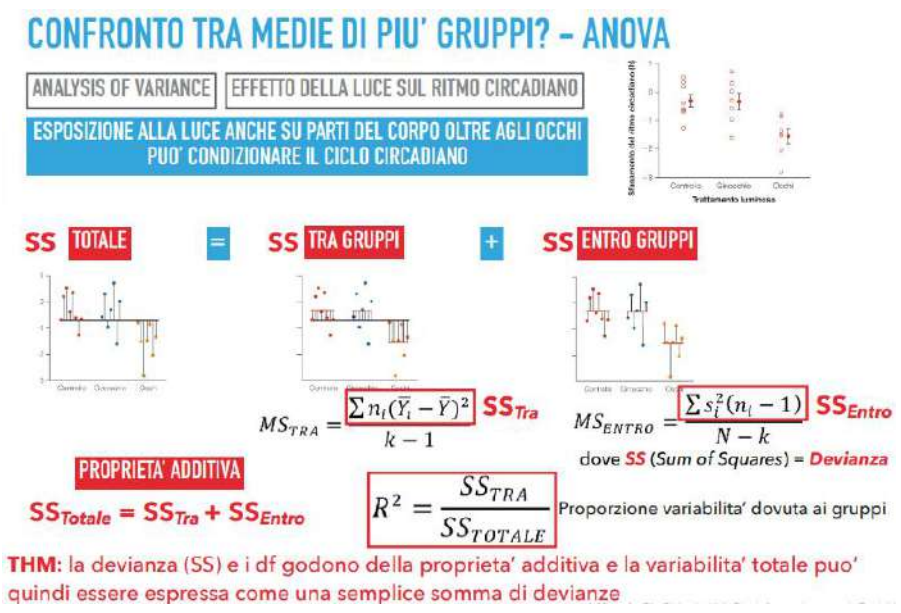
Da qui si ricava **F, che è il RAPPORTO TRA LE VARIANZE**: assumendo vera l'ipotesi nulla che le medie delle popolazioni di tutti i gruppi siano uguali, il rapporto F:

$$F = \frac{MS_{gruppi}}{MS_{errore}}$$

deve essere VICINO A 1. Se invece è falsa l'ipotesi nulla, il rapporto sarà MAGGIORE DI 1. Es. nei dati del ritmo circadiano abbiamo F=7,29. Come facciamo a capire se è significativo? Usiamo il P value! **La distribuzione nulla per la statistica F è la DISTRIBUZIONE F**, che ha una coppia di gradi di libertà, uno per il numeratore (k-1) e uno per il denominatore (N-k). È importante tener conto che prima si cerca quella del numeratore e POI quella del denominatore. **Ora, dalle tavole si può vedere che per alpha=0,05 con al numeratore (COLONNE, TRA) 2gl e al denominatore (RIGHE, ENTRO) 19gl abbiamo F=3,52. Quindi SI RIFIUTA L'IPOTESI NULLA**.

Questi sono tutti p value quelli delle tabelle!!!!

Un valore importante è **R^2,** che viene usato per valutare il **CONTRIBUTO DELLE DIFFERENZE TRA I GRUPPI ALLA VARIABILITA' DEI DATI**.



Questa grandezza si basa sul fatto che:

SStot=SStra+SSentro (LA SOMMA NON VALE ANCHE PER I MEAN SQUARE)

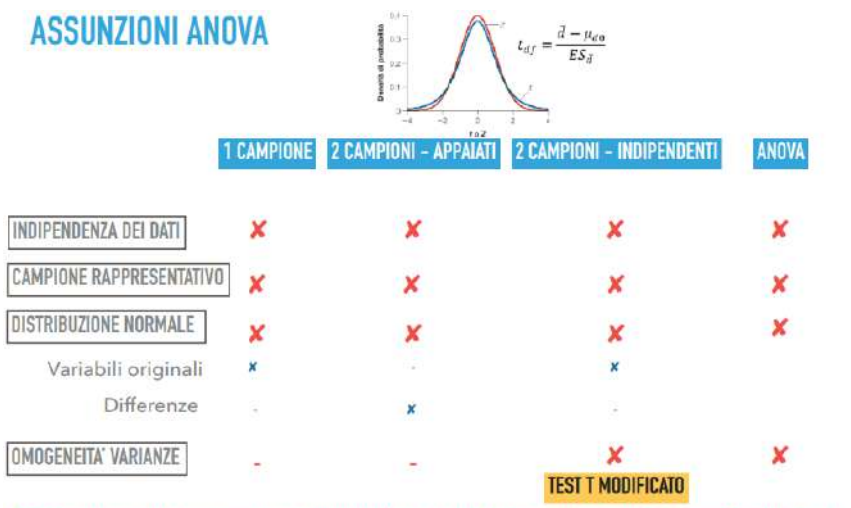
E quindi

$$R^2 = \frac{SS_{gruppi}}{SS_{totale}}$$

Dice quanto è più stretta la dispersione delle misure intorno alle medie dei gruppi rispetto a quella intorno alla media generale. Ha valori compresi tra 1 e 0. Se è prossimo a 0, vuol dire che le medie dei gruppi sono tutte molto simili, e la maggior parte della variabilità è all'interno dei gruppi. Quando invece è vicino a 1 resta poca variazione in Y dopo che si sono considerate le differenze tra le medie dei gruppi: quindi la maggior parte della variabilità di Y è SPIEGATA DALLA VARIABILE ESPLICATIVA.

Per esempio nel ciclo circadiano abiamo 0,43. Questo vuol dire che il 43% della somma dei quadrati totale nel valore dello sfasamento è spiegato da differenze tra individui sottoposti a diversi trattamenti, mentre il 57% dalla differenza tra gli individui, che è un semplice ERRORE.

Quindi R^2 misura la FRAZIONE DELLE VARIABILITA' DI Y CHE è SPIEGATA DALLE DIFFERENZE TRA I GRUPPI (e non dalla variabilità naturale)



ASSUNZIONI ANOVA

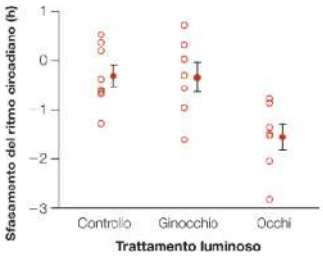
- Le misure di goni gruppo sono un campione casuale della popolazione
- La variabile è distribuita normalmente in ogni popolazione
- La varianza è la stessa in tutte k le popolazioni

Qui si vedono tutti i modi per controllare che le assunzioni siano rispettate.

CONFRONTO TRA MEDIE DI PIU' GRUPPI? – ANOVA NPAR

H DI KRUSKAL-WALLIS | COME IL MW MA PER PIU' GRUPPI

ORDINARE
RANGHIZZARE
SOMMARE RANGHI PER GRUPPO
STIMARE LA STATISTICA H
CONFRONTARE LA STAT CON QUELLA ATTESA SOTTO H_0



$$H = \frac{12}{N(N+1)}\sum\frac{R_i^2}{n_i} - 3(N+1)$$

H approssima la distribuzione del X^2 con df = k-1

THM: MW puo' essere esteso a 3 gruppi usando il test di Kruskal-Wallis

ALTERNATIVE NON PARAMETRICHE ALL'ANOVA

Come nel test MW ma PER PIU' GRUPPI.

Il test di cui parliamo è il **TEST DI KRUSTAL WALLIS**.

Anche qui, come il test U di MW bisogna avere che la distribuzione della variabile abbia la stessa forma in ogni popolazione.

La distribuzione campionaria di H è molto simile a quella del CHI QUADRO con k-1 gradi di libertà, dove k è sempre il numero di gruppi. L'ipotesi nulla viene rifiutata se H è superiore o uguale al valore critico ottenuto dalla distribuzione chi quadro.

Ha una potenza limitata se le dimensioni del campione sono molto ridotte. Quindi è **da preferire l'ANOVA**.

TABELLA ANOVA (F)

Cumulative Probabilities = 0,95

Upper-Tail Areas = 0.05

Numerator, df_1

Denominator, df_2	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞	Denominator, df_2
1	161.40	199.50	215.70	224.60	230.20	234.00	236.80	238.90	240.50	241.90	243.90	245.90	248.00	249.10	250.10	251.10	252.20	253.30	254.30	1
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50	2
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53	3
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63	4
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36	5
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67	6
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23	7
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93	8
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71	9
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54	10
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40	11
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30	12
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21	13
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13	14
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07	15
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01	16
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96	17
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92	18
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88	19
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84	20
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81	21
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.91	1.89	1.84	1.78	22
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76	23
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73	24
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71	25
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69	26
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67	27
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65	28
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64	29
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62	30
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51	40
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39	60
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25	120
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00	∞
Denominator, df_2	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**15**	**20**	**24**	**30**	**40**	**60**	**120**	**∞**	**Denominator, df_2**

ANOVA PIANIFICATA

L'analisi della varianza di solito è l'inizio del confronto tra le medie di più di due gruppi. Poi ci si deve rispondere ad altre domande, come: **QUALI MEDIE SONO DIVERSE? Qual è l'entità della differenza tra le medie**? L'ANOVA da solo non risponde a nessuna di queste domande. Per determinare quali medie sono diverse e di quanto di solito si usano due approcci:

il **confronto PIANIFICATO è un confronto tra medie ritenute di INTERESSE CRUCIALE NELLO STUDIO, quindi PRIMA di raccogliere i dati. Quindi questo confronto è GIUSTIFICATO A PRIORI**!

Per esempio, lo è la previsione di una teoria o di uno studio precedente. Un confronto **NON PIANIFICATO** invece è uno dei tanti possibili confronti multipli tra tutte le coppie di medie, per identificare A POSTERIORI le medie diverse. In questo caso è necessario **PROTEGGERSI DALL'AUMENTO DEGLI ERRORI DI TIPO 1**.

Un buon esempio di un confronto PIANIFICATO è l'esperimento sul ciclo circadiano. L'esperimento propone il **confronto tra lo sfasamento medio della produzione di melatonina tra i gruppi che hanno ricevuto luce al ginocchio e non hanno ricevuto luce**. quindi PER QUESTO CONFRONTO possiamo **procedere in modo PIANIFICATO**, chiedendoci quando è grande la differenza tra il gruppo sottoposto al trattamento alle ginocchia e quello di controllo e se la distanza sia o meno significativa.

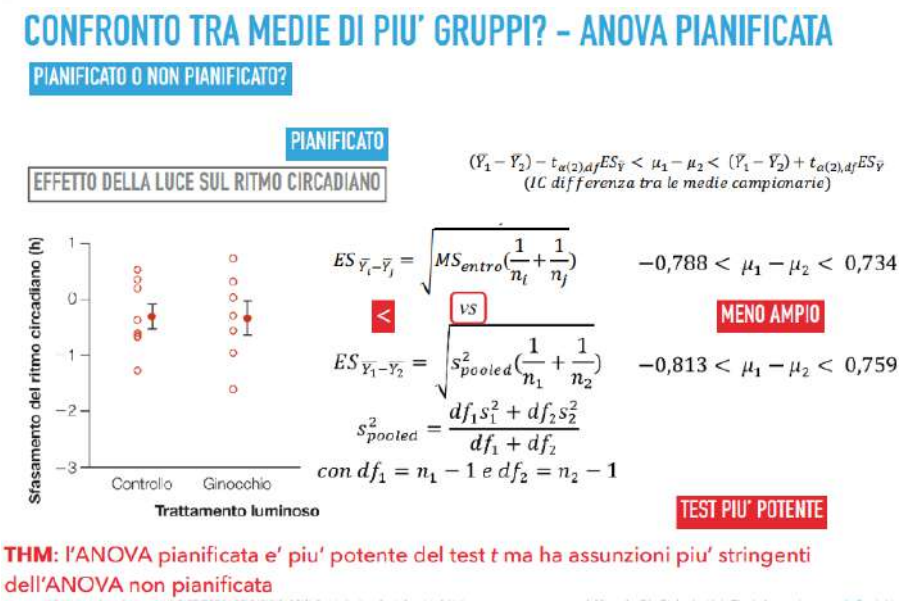
Il metodo per fare un confronto pianificato è molto simile al t test tra due campioni. Solo l'ERRORE STANDARD viene CALCOLATO IN MODO DIVERSO: nel confronto pianificato si usa la varianza campionaria comune basata su tutti e k i gruppi e NON quella basata solo sui due gruppi. In questo modo AUMENTA LA PRECISIONE E LA POTENZA.

Per esempio si può voler analizzare la differenza tra le medie dei gruppi di trattamento al ginocchio e di controllo.

$$\bar{Y}_2 - \bar{Y}_1 = (-0{,}336) - (-0{,}309) = -0{,}027 \text{ h}$$

L'errore standard per questa differenza è dato da

$$ES = \sqrt{MS_{error}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$



Con N-k gradi di libertà. Si usa la **DISTRIBUZIONE t** e si trova:

$$-0{,}788 < \mu_2 - \mu_1 < 0{,}734$$

Se usassimo solo il confronto a due campioni, otterremmo UN INTERVALLO DI CONFIDENZA PIU' LARGO.

Quindi questo test è leggermente più potente.

CONFRONTI NON PIANIFICATI

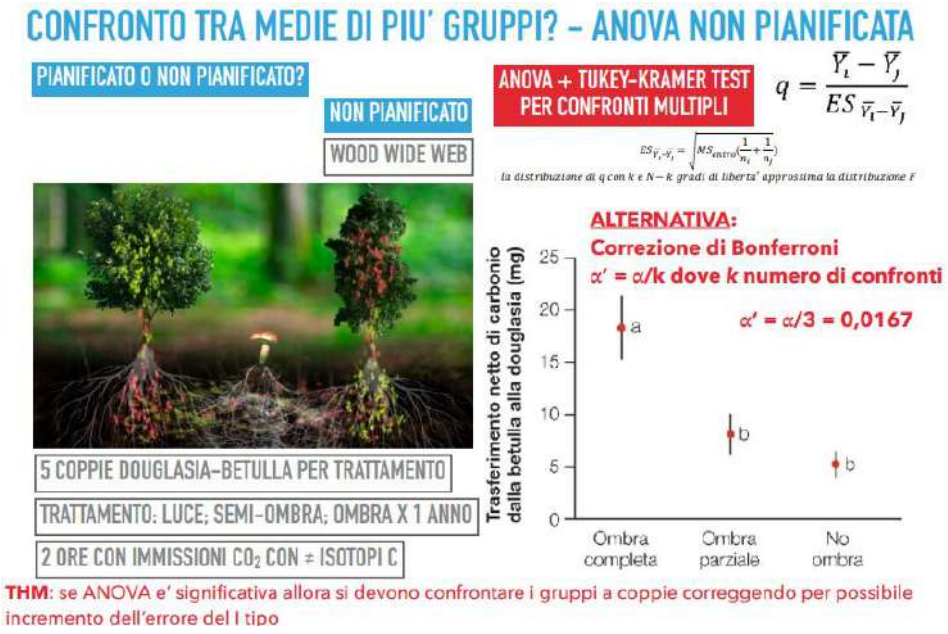
La formula usata per i confronti pianificati non vale per quelli non pianificati. **Richiedono infatti degli AGGIUSTAMENTI** per **IMPEDIRE L'AUMENTO DEL TASSO DEI FALSI POSITIVI**.

Il numero di possibili confronti che coinvolgono k gruppi è potenzialmente molto alto.

Bisogna VALUTARE TUTTE LE COPPIE DI MEDIE per trovare quali gruppi sono diversi dagli altri. Il test più usato è il **TEST TUKEY KRAMER, che assume che SIA GIA' STATO ESEGUITO UN ANOVA**.

Esempio: associazioni simbiontiche delle piante. Studio del flusso di carbonio tra le piante di betulla e di abete per vedere se dipendesse dall'ombra sulle piante. Gli alberi in ombra infatti potrebbero estrarre più carbonio attraverso le micorrize rispetto a quelli in pieno sole. La maggior parte del trasferimento di carbonio è avvenuto dalla betulla all'abete. I trattamenti con diversi livelli di ombra hanno determinato differenze nell'incremento medio di carbonio da parte delle piantine di abete, e si vede da F alto e basso P. ma **QUALI MEDIE SONO DIVERSE TRA LORO? Ogni media è diversa dalle altre? Oppure quali sono le coppie di medie diverse**?
Questo METODO TK opera come una serie di test t per due campioni, ma USA UN VALORE CRITICO MAGGIORE PER LIMITARE IL TASSO DI ERRORE DI TIPO 1. **In questo modo, la possibilità di commettere almeno un errore di tipo 1 nel corso della comparazione di tutte le coppie di medie NON è maggiore di alpha**!
Come si fa:

- ORDINARE LE MEDIE, in ordine crescente
- CONFRONTARE OGNI COPPIA DI MEDIE, in modo che si verifichi:
  IPOTESI NULLA: μ1 = μ2
  IPOTESI ALTERNATIVA: μ1 != μ2
- La STATISTICA TEST è la Q della SLIDE: si basa sul **mean square degli errori**.
- Si confronta questa statistica test con la DISTRIBUZIONE Q avente k e N-k gradi di libertà. **I VALORI CRITICI SONO QUELLI DELLA STATISTICA F**!
- In questo caso si usano dei simboli per indicare il risultato del test, come si vede dalla slide: i punti e le barre di errore indicano le medie + o – l'errore standard. **A due medie è assegnata una lettera diversa se sono significativamente diverse, se invece il simbolo è uguale vuol dire che NON sono significativamente diverse**.



Si parla di **CORREZIONE DI BONFERRONI, per cui alpha'=alpha/k**, per cui **CAMBIA L'ALFA DI RIFERIMENTO grazie al fatto che si usi il test q**. Così la probabilità si commettere errore di tipo 1 NON è maggiore di alpha, ma arriva ad alpha stesso.

Confronti MULTIPLI: in biomed si fanno sempre confronti di questo tipo usando i CONTROLLI: cioè non si confrontano i gruppi tra loro, ma il gruppo con il controllo

CONFRONTI MULTIPLI PIANIFICATI CON CONTROLLO

Per esempio: 4 trattamenti e un controllo — **4 CONFRONTI**

TEST DI DUNNETT — Formulazione simile al Tukey-Kramer

$$Q_{(\alpha,p,\nu)} = \frac{\overline{X}_c - \overline{X}_i}{\sqrt{s_e^2 \cdot \left(\frac{1}{n_c} + \frac{1}{n_i}\right)}} \qquad s^2 = \frac{\sum_{i=0}^{p} \sum_{j=1}^{N_i} (X_{ij} - \bar{X}_i)^2}{n}$$

dove

- **c = gruppo di controllo o placebo**
- **i = trattamento in oggetto**, per il quale si effettua il confronto con il controllo
- **p = numero di confronti** (uguale al numero di gruppi (k) meno 1).

THM: Il test di Dunnett si usa per confrontare i gruppi di trattamenti con un gruppo di controllo ma non i trattamenti tra loro e quindi riduce il numero di confronti ed è per questo più potente

Questo implica l'uso del test DUNNET, che è quello che si deve usare se si ha un **gruppo di controllo di riferimento e si FANNO CONFRONTI MULTIPLI RISPETTO AL SINGOLO CONTROLLO E NON TRA GRUPPI SPERIMENTALI, MA IL SINGOLO GRUPPO SPERIMENTALE RISPETTO AL CONTROLLO**.
È molto utile perché cambia il numero di confronti.
Sarebbe bello saperlo nel compito, scrivendo avrei dovuto usare il Dunnet, ma uso il TUKEY perché è più facile da fare a mano.

CONFRONTI MULTIPLI PIANIFICATI CON CONTROLLO

TEST DI DUNNETT

$$Q_{(\alpha,p,\nu)} = \frac{\overline{X}_c - \overline{X}_i}{\sqrt{s_e^2 \cdot \left(\frac{1}{n_c} + \frac{1}{n_i}\right)}}$$

N – K (GRUPPI)

Valori critici del Q per il test di Dunnett bilaterale

- k = numero di medie a confronto (compreso il controllo)
- ν = gradi di libertà della varianza d'errore

$\alpha_T = 0.05$ bilaterale

NUMERO k DI MEDIE A CONFRONTO

ν	2	3	4	5	6	7	8	9	10
8	2,31	2,67	2,88	3,02	3,13	3,22	3,29	3,35	3,41
9	2,26	2,61	2,81	2,95	3,05	3,14	3,20	3,26	3,32
10	2,23	2,57	2,76	2,89	2,99	3,07	3,14	3,19	3,24
11	2,20	2,53	2,72	2,84	2,94	3,02	3,08	3,14	3,19
12	2,18	2,50	2,68	2,81	2,90	2,98	3,04	3,09	3,14
13	2,16	2,48	2,65	2,78	2,87	2,94	3,00	3,06	3,10
14	2,14	2,46	2,63	2,75	2,84	2,91	2,97	3,02	3,07
15	2,13	2,44	2,61	2,73	2,82	2,89	2,95	3,00	3,04
16	2,12	2,42	2,59	2,71	2,80	2,87	2,92	2,97	3,02
17	2,11	2,41	2,58	2,69	2,78	2,85	2,90	2,95	3,00
18	2,10	2,40	2,56	2,68	2,76	2,83	2,89	2,94	2,98
19	2,09	2,39	2,55	2,66	2,75	2,81	2,87	2,92	2,96
20	2,09	2,38	2,54	2,65	2,73	2,80	2,86	2,90	2,95
24	2,06	2,35	2,51	2,61	2,70	2,76	2,81	2,86	2,90
30	2,04	2,32	2,47	2,58	2,66	2,72	2,77	2,82	2,86
40	2,02	2,29	2,44	2,54	2,62	2,68	2,73	2,77	2,81
60	2,00	2,27	2,41	2,51	2,58	2,64	2,69	2,73	2,77
120	1,98	2,24	2,38	2,47	2,55	2,60	2,65	2,69	2,73
∞	1,96	2,21	2,35	2,44	2,51	2,57	2,61	2,65	2,69

THM: semplice e potente

A Massolo, Dip Biologia, Univ Pisa | alessandro.massolo@unipi.it

La correzione di Bonferroni è principalmente progettata per ridurre il rischio di errore di tipo I, che si verifica quando si rifiuta erroneamente un'ipotesi nulla vera. Tuttavia, è importante notare che l'uso della correzione di Bonferroni può aumentare il rischio di errore di tipo II, che si verifica quando non si rifiuta un'ipotesi nulla falsa.

Quando si applica la correzione di Bonferroni dopo la procedura di Tukey-Kramer, è importante essere consapevoli di questo trade-off tra errori di tipo I e tipo II. La correzione di Bonferroni riduce la probabilità di commettere errori di tipo I, ma può rendere più difficile rilevare differenze vere tra i gruppi, aumentando il rischio di errori di tipo II.

La scelta di utilizzare la correzione di Bonferroni dovrebbe essere guidata dalla considerazione della situazione specifica e degli obiettivi dell'analisi. Se la tua priorità principale è controllare il rischio di scoprire erroneamente differenze tra i gruppi, la correzione di Bonferroni può essere appropriata. Tuttavia, se la sensibilità nel rilevare

effettive differenze tra i gruppi è più importante, potresti optare per un livello di significatività meno conservativo o esplorare altre strategie di correzione.

In generale, la scelta di applicare la correzione di Bonferroni dipende dalle tue esigenze specifiche e dalle considerazioni sulla gestione degli errori di tipo I e II nel contesto della tua analisi statistica

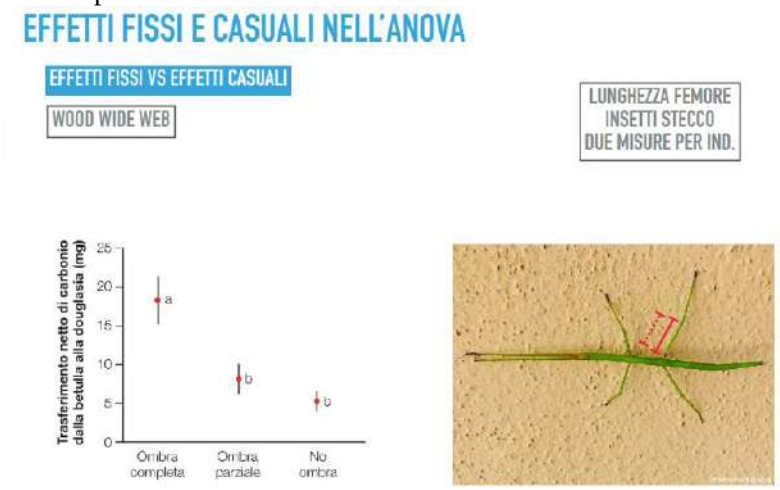
EFFETTI FISSI E CASUALI DELL'ANOVA

Finora abbiamo visto gruppi FISSI, cioè studi in cui le diverse categorie della variabile esplicativa sono predeterminate e ripetibili.

Esempi di EFFETTI FISSI: trattamenti medici alternativi in una sperimentazione clinica, dosi prefissate di una tossina, diverse altezze rispetto alla bassa marea nella zona intertidale, gruppi suddivisi per sesso o età.

Ogni conclusione che si raggiunge riguardo alle differenze tra gruppi fissi **è VALIDA SOLO PER QUESTI GRUPPI**! Per esempio, se per una certa variabile di risposta si trovasse una differenza tra due trattamenti farmacologici, **NON SI POTREBBERO GENERALIZZARE I RISULTATI AD ALTRI FARMACI**!

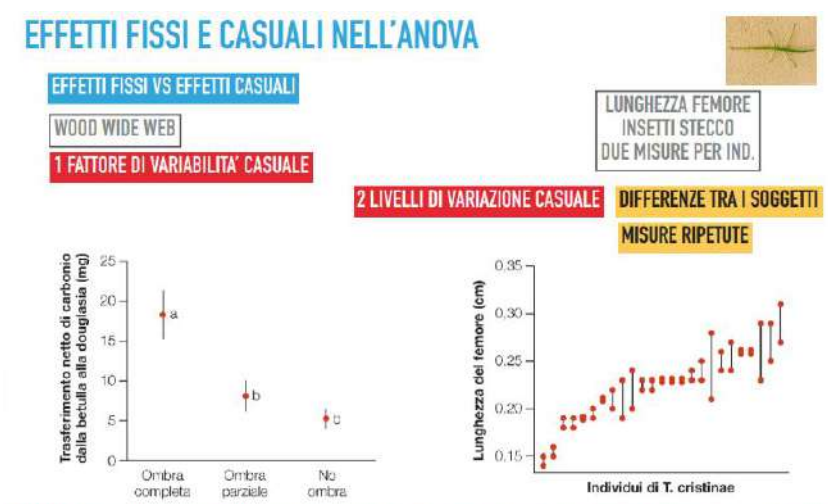
Esiste però un SECONDO TIPO DI ANOVA, in cui i gruppi NON sono fissi ma sono SCELTI CASUALMENTE! Questo vuol dire che i gruppi NON sono predeterminati, ma **CAMPIONATI CASUALMENTE DA UNA POPOLAZIONE MOLTO PIU' GRANDE DI GRUPPI**. Questo ANOVA applicato a gruppi casuali è detto a EFFETTI CASUALI. Esempi: la famiglia in uno studio sulla somiglianza tra parenti in relazione al QI; l'individuo, in uno studio sull'errore di misura condotto facendo misurazioni diverse per diversi individui.



Esempio: vogliamo valutare se le famiglie differiscono l'una dall'altra per i valori di QI medi dei bambini. Dobbiamo quindi partire da un campione casuale di famiglie rispetto ad una popolazione di famiglie, e la **<famiglia> verrebbe usata come VARIABILE di GRUPPO in un ANOVA**, e le repliche sarebbero I BAMBINI ALL'INTERNO DI OGNI FAMIGLIA. In uno studio successivo uguale sulla stessa popolazione, NON avrebbe senso prendere le stesse famiglie, perché **IL TARGET DELLO STUDIO è LA POPOLAZIONE, NON LE SPECIFICHE FAMIGLIE ESTRATTE CASUALMENTE. I RISULTATI PERO' SI GENERALIZZANO ALLA POPOLAZIONE INTERA**.

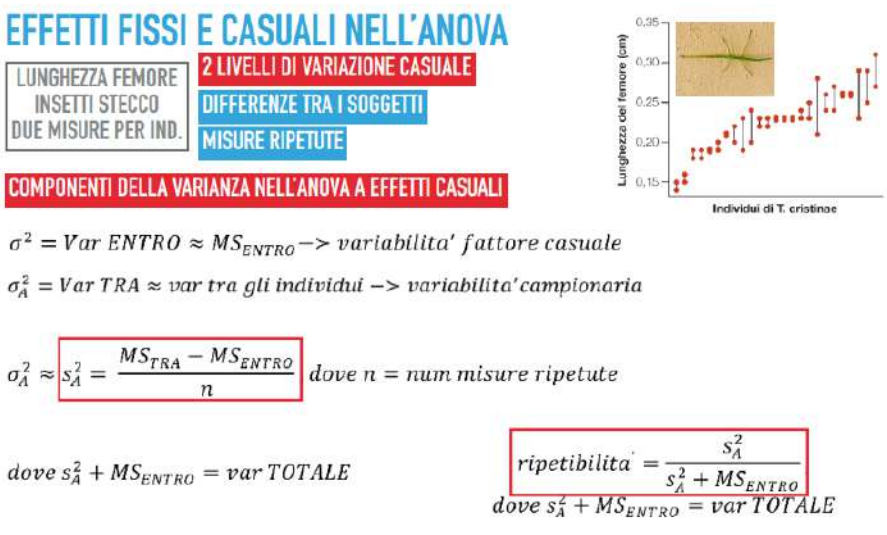
Una variabile esplicativa è un effetto fisso se i gruppi sono predefiniti e sono di interesse specifico. Una variabile esplicativa è un effetto casuale se i gruppi sono campionati casualmente da una popolazione di gruppi possibili.

Esempio ANOVA a effetti casuali: può essere usato per STIMARE LE COMPONENTI DELLA VARIANZA (aka la varianza entro e tra i gruppi):



esempio: insetto stecco. Misurazioni di vari parametri del corpo da immagini al computer. Essendo interessati all'errore di misura sono state scattate due foto per ogni animale. la seconda misura era spesso diversa dalla prima. Quanto era grande l'errore di misura rispetto alla variazione reale tra i caratteri degli individui?
dai risultati dell'ANOVA c'è scritto che tra i gruppi sono presenti differenze significative. Nonostante l'errore di misura, sono ancora presenti differenze identificabili nelle lunghezze del femore tra i diversi esemplari. **Questo vuol dire che tra i singoli INSETTI NELLA POPOLAZIONE ci SONO DIFFERENZE**!

Questo era un ANOVA A UN FATTORE A EFFETTI CASUALI, e differisce dall'ANOVA a effetti dissi perché nel primo caso vengono **considerati due livelli di variazione casuale rispetto alla variabile di risposta, cioè la VARIAZIONE ENTRO I GRUPPI e TRA I GRUPPI**. Per esempio, nello studio dell'insetto stecco, il primo livello (ENTRO i gruppi) è la **VARIANZA TRA LE MISURE RIPETUTE SULLO STESSO INDIVIDUO, che è CAUSATA SOLO DA ERRORI DI MISURAZIONI**. L'ANOVA assume che l'errore di misura sia lo stesso per ogni insetto misurato, a parte l'effetto del caso. Per valutare invece la varianza TRA i gruppi si deve assumere che OGNI GRUPPO HA LA SUA MEDIA. e l'ANOVA a effetti casuali assume che le medie dei gruppi abbiano una distribuzione di probabilità NORMALE nella popolazione.
I DUE PARAMETRI CHE SI OTTENGONO, cioè la VARIANZA della lunghezza media tra gli individui della popolazione e la VARIANZA ENTRO I GRUPPI DELLA POPOLAZIONE sono le COMPONENTI DELLA VARIANZA e descrivono tutta la varianza della variabile di risposta.



FORMULA DELLA RIPETIBILITA' NELLA SLIDE: misura della similarità complessiva delle misure ripetute nello stesso gruppo.
Una ripetibilità VICINA A 1 indica che le misurazioni ripetute effettuate nello stesso gruppo forniscono OGNI VOLTA LO STESSO RISULTATO. Esempio, se è 0,75 vuol dire che una percentuale del 75% della varianza totale della lunghezza del femore degli insetti stecco è dovuta alle differenze tra i singoli insetti. Il restante 25% è dovuto invece agli errori di misura.

LAVORO SCIENTIFICO

Come valutare un lavoro scientifico? Hanno veramente scritto nel modo migliore in cui potevano?
PER FARE L'ESAME: CARICARE ONLINE IL FILE CON LA CRITICA

Pubblicazione accettabile: criteri
Un lavoro ascientifico è una sintesi formale di un lavoro di ricerca che vuole rispondere a delle domande e confutare delle ipotesi
STRUTTURA
Un lavoro scientifico deve rispondere a determinate domande, tra cui**: qual è il problema, come è stato affrontato il problema** (materiali e metodi), **cosa hai trovato** (risultati) **e discussione, cioè cosa hai ricavato dai risultati,** che si deve inserire nella letteratura corrente.
La struttura è detta IMRD: introduction, materials, results, discussion. Ci sono delle riviste che cambiano l'ordine dei capitoli ma la maggior parte degli articoli hanno la struttura in una data sequenza. La prima cosa che si vede in un articolo è il TITOLO, poi abstract, e poi IMRD. Una cosa imperante sono anche le **LIMITAZIONI**.

Il titolo non deve essere una frase lunghissima: informativa e succinta. L'abstract è una sezione che identifica il contenuto di base: cosa si trova, perché ha senso quello che si trova, è un riassunto che deve sintetizzare quello che si è fatto, cosa si è fatto e il contenuto. Titolo e abstract sono fondamentali quando si presenta la tesi e un progetto di ricerca, perché spesso sono le uniche cose che si leggono. L'introduzione si scrive per ultima.
Materiali a metodi: si deve prendere un testo complesso e renderlo di facile lettura. È bene fornire tutte le **informazioni necessarie per replicare quello che abbiamo fatto fornendo la struttura degli obiettivi**.
Importante dire perché sono state usate delle procedure, mettere l'articolo di riferimento in cui le spiegano e se le abbiamo modificate perché e come le abbiamo modificate. La struttura di materiali e metodi deve essere replicata nei risultati. Poi la discussione, che di nuovo: prende il primo risultato e lo discute. Per ogni risultato rilevante bisogna interpretare il senso biologico e dire il meccanismo che sottende quello che abbiamo scritto e vedere cosa dice la letteratura in casi simili. Poi nella conclusione una sintesi di tutto quello che abbiamo fatto e dedotto.
Tutte le conclusioni sono rilevanti, quindi è importante per ogni conclusione dire quali sono i limiti.

Cosa merita di essere pubblicato?
Articoli che dimostrano, spiegano, introducono. Review.
Di solito pubblicare costa (se non hai soldi e sei fortissimo potresti essere esente dal pagamento). Bisogna scegliere l'audience, e in questo modo posso anche capire in che modo scriverlo.
Insomma, anche per pubblicare servono soldi.
Ogni articolo ha un suo formato. in base al tipo di articolo la lunghezza delle pagine cambia. Ci sono le short com, che sono piccoli lavori che non meritano un paper intero ma è un peccato non pubblicarli. In alcune riviste non c'è il controllo editoriale.
Devo poi scrivere una cover letter in cui dico al revisore perché dovrebbe pubblicare il mio lavoro. spesso capita che un singolo revisore non sia abbastanza, ma un lavoro viene mandato anche a 5 revisori esperti del mestiere. Alcune riviste hanno un sistema per cui anche l'autore è anonimo per il reviewer. Dopo questa fase, se l'articolo viene accettato c'è la valutazione finale e i commenti del reviewer. I commenti sono utili per migliorare il lavoro, anche se l'EDITORE non accetta l'articolo (per ora lo ha accettato solo il reviewer)

SUBMISSION OUTCOME
Cosa ti possono dire del tuo lavoro?
- Rifiutato: di solito ti consigliano una rivista a cui mandarlo, che è spesso una sottorivista.
- A volte il risultato dice di revisionare di nuovo e risottomettere.
- A volte le revisioni che si devono fare sono minori. Se devi fare delle revisioni vuole comunque dire che non sei stato rifiutato e puoi essere accettato

A questo punto c'è la REVISIONE. Ad ogni commento del revisore si deve scrivere una risposta. Spesso comunque si deve risottomettere il lavoro. Molto raramente c'è il pass diretto, anche se di solito almeno le minor revisions ci sono.

Come trovare un articolo scientifico?

- WEB OF SCIENCE
- SCOPUS
- PUBMED

Ci sono gli articoli in riviste indicizzate, cioè riviste che appartengono a degli indici e hanno controllo di qualità
C'è anche SCHOLAR, che però controlla la letteratura grigia.
PER L'ESAME BISOGNA TROVARE LA RIVISTA IN UNA DI QUESTE TRE SOPRA

IF: è 1 se ho 100 articoli e mi citano 200 volte, vuol dire che ho 200 citazioni su 100 articoli, quindi ho un IF di 2.
Limiti del IF: alcune discipline lo hanno più facilmente alzabile di altre.
Spesso non è tanto questo numerello, ma il numerello rispetto a quello del settore disciplinare in cui è, perché ci sono discipline che lo hanno più alto di solito. Quindi 1) vedo in che rivista è pubblicato l'articolo, 2) controllo l'IF di quella rivista e 3) lo confronto con quello delle altre riviste dello stesso settore.

CRITERI GENERALI PER REVISIONARE UNA PUBBLICAZIONE
Per il lavoro che dobbiamo fare per l'esame. Chiunque è revisore di chiunque.
Quando si revisiona un articolo, questo aiuta a capire dove si commettono più frequentemente degli errori. Ci sono diverse tipologie di articoli che possono essere sottomessi.
MEGLIO SCEGLIERE UN **RESEARCH PAPER, con almeno un po' di statistica sopra**.
Criteri: ci servono per completare il compito del Massolo e riempire delle caselle in un file excel: si valuta il paper sulla base di pro e contro di un articolo fatto bene, che indichiamo di seguito.

TITOLO
Tabella con diverse sezioni e frasi per capire se le singole sezioni sono o meno soddisfacenti. Il titolo è soddisfacente quando contiene il minor numero possibile di parole e contenga magari ciò che è stato ottenuto. Non deve essere una frase di senso compiuto ma può esserlo e può dare una informazione ad un risultato importante, soprattutto se è un bel risultato. Non è soddisfacente se è troppo lungo, con troppe parole lunghe, con la struttura grammaticale non fatta bene… con il titolo cono comunque aspetti un pochino più soggettivi più che oggettivi come la parte statistica
Quando si sceglie che titolo dare all'articolo scientifico ci si deve immaginare che contenga delle parole chiave che aiutano chi sta cercando l'argomento in modo facile.
ABSTRACT
Deve essere conciso e molto chiaro, includere almeno una frase relativa a tutte le sezioni dell'articolo. Quello che è insoddisfacente è che l'abstract non ci sia e che sia focalizzato poco sulle tematiche affrontate e le conclusioni dell'articolo. A volte infatti si legge l'abstract e non si capisce cosa è stato fatto.
INTRODUZIONE
Deve contenere il razionale dello studio. Per giudicare bisogna essere esperti di quello che si sta leggendo. Quindi bisogna fare un minimo di ricerca bibliografica per capire se nell'introduzione e poi nella discussione sono stati citati adeguatamente gli studi più importanti del settore. I riferimenti bibliografici sono adeguati? Alla fine dell'introduzione deve essere descritto in modo chiaro il razionale.
METODI
La descrizione deve essere appropriata. Devono essere descritti bene, sia per il lavoro sul campo, sia per il sistema di studio, il numero di repliche, le dimensioni delle unità sperimentali, le ANALISI STATISTICHE, che tipo sono e come sono state fatte.
Questo perché CHIUNQUE DEVE POTER REPLICARE L'ESPERIMENTO!
RISULTATI

Si segue lo stesso ordine con cui si sono proposte le ipotesi da verificare. Le tabelle e le figure devono essere adeguate alle domande che ci si è posti, e le figure in particolare devono essere adeguate a risultati che mettano in evidenza l'analisi statistica.
È importante che la figura sia fatta bene (BOX PLOT, NON BAR PLOT). Si capisce cosa hanno fatto? Qual era l'n campionario?
Le figure non devono essere troppe. Bisogna che ci siano solo il numero di box plots a sufficienza, non ridondanti e solo quelli assolutamente rilevanti. Le differenze tra i gruppi devono essere significative.

DISCUSSIONE

Deve essere una discussione critica dei nostri risultati, a partire dalle nostre ipotesi. Si deve seguire la stessa scaletta logica di materiali e metodi. È facile che quando si scrivono i risultati si ceda alla tentazione di andare oltre a quello che i risultati dicono. Si deve rimanere entro i limiti. Normalmente nella discussione si fa riferimento a studi che hanno fatto parte del lavoro, come sistemi di studio che hanno fatto una cosa. Na cosa sbagliata è generalizzare troppo quello che è stato fatto: per esempio, se ho fatto una cosa, dico che c'è uno studio che dice che con delle determinate condizioni lo dovevo fare così. Non devi essere neanche troppo critico.
Importante il fatto che **LA NON SIGNIFICATIVITA' RIMANE TALE. Spesso il fatto che una cosa non è significativa vuol dire che una differenza non è stata rilevata. Se il p value è comunque elevato anche se non significativo, si può parlare di una TENDENZA ALLA SIGNIFICATIVITA'**. Non si può dire che un risultato è significativo però.

CONCLUSIONI

Parte finale della discussione. Bisogna dire solo cose vere

CITAZIONI

Si deve citare tutti gli articoli importanti. Attenzione ai dettagli: non fa piacere vedere che gli autori non hanno seguito le indicazioni specifiche della rivista su come riportare la letteratura. Esistono alcune riviste che permettono di sottomettere l'articolo scegliendo il formato delle citazioni, l'importante è che ci sia coerenza. Quello che non deve accadere è che alcuni articoli siano citati in un modo, altri in un altro. La cosa importante è che i revisori non citino l'articolo di riferimento.
ALLE CITAZIONI FACCIAMO ATTENZIONE DURANTE LA TESI MA NON PER IL LAVORO DI MASSOLO.

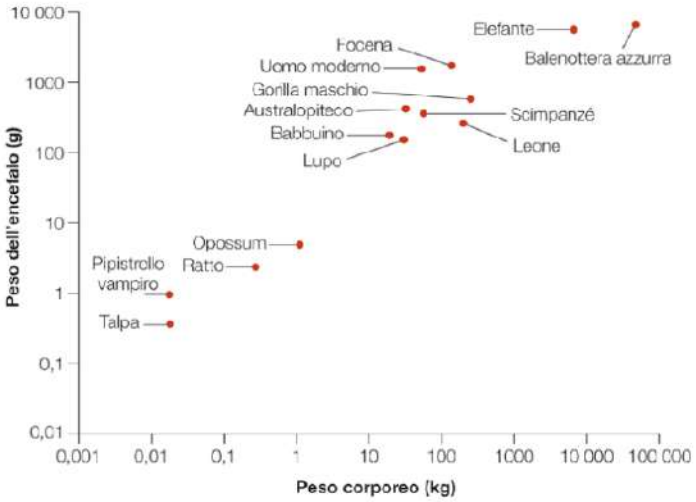
La cosa che dobbiamo fare è dare una risposta a ciascuna delle domande del file excel online su elearning. Poi alleghiamo il PDF su elearning in una cartella apposita. Le domande più importanti sono quelle in grassetto, quelle più inerenti alla discussione in classe. La statistica è il COLLANTE. La discussione si fa a seguito delle ANALISI STATISTICHE.
ESERCIZIO SUL PORTALE ELEARNING: SI CARICA IL PDF DELL'ARTICOLO E IL FOGLIO EXCEL!

Esempio articolo

Rivista sua, è una NOTA, cioè un articolo piccolino con alcuni dati che possono essere interessanti. Il titolo DICE IL RISULTATO.
Può capitare di trovare anche un ABSTRAT GRAFICO, oltre a quello scritto, che riassume quello che è stato fatto in modo grafico.
Risultati: i grafici fatti a istogramma non sono belli secondo Massolo, perché c'è un istogramma altissimo e gli altri molto bassi, e soprattutto **NON HA I PALLINI SU COME SONO DISTRIBUITI I CAMPIONI**.
Il **box plot non ne ha bisogno, per questo è il top, ma SI FA SOLO SE ABBIAMO ALMENO 7 CAMPIONI, SE NO NON HA SENSO**.
Se si ottiene un risultato non significativo NON vuol dire che NON c'è effetto della manipolazione, solo CON UN CERTO SETTING SPERIMENTALE NON è STATO OTTENUTO, ma potrebbe esserci se si usano più campioni o le ricerche vengono fatte in modo diverso.
LA PRIMA COSA CHE DEVO FARE è VEDERE L'IMPACT FACTOR.

REGRESSIONE LINEARE

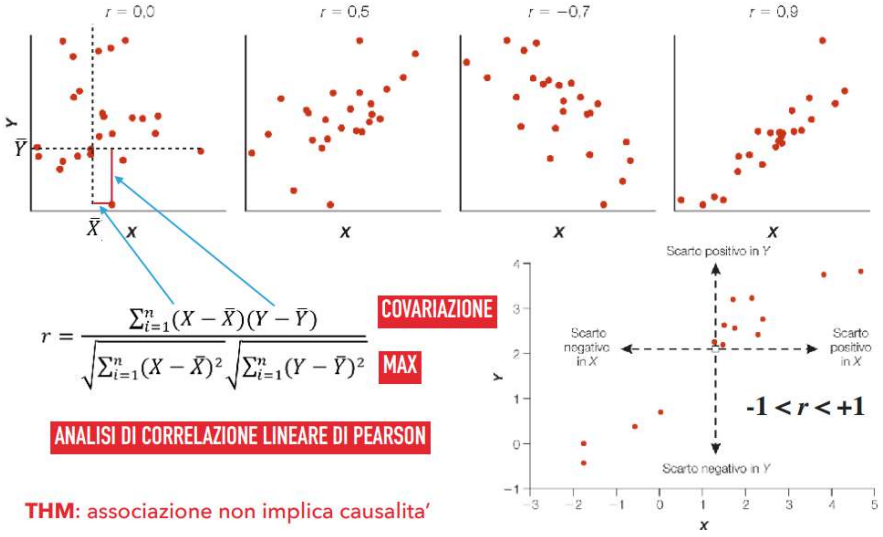
L'ASSOCIAZIONE TRA VARIABILI NUMERICHE



Specie	Body (kg)	Brain (g)
talpa	0.04	0.70
pipistrello vampiro	0.04	1.00
ratto	0.45	4.00
opossum	1.00	7.00
babbuino	14.00	150.00
lupo	15.00	145.00
australopiteco	70.00	500.00
-	-	-
-	-	-
-	-	-
-	-	-
Balenottera azzurra	80,000.00	7,500.00

THM: misura associazione tra 2 variabili diverse (NON DATI APPAIATI)

L'ASSOCIAZIONE TRA VARIABILI NUMERICHE



THM: associazione non implica causalita'

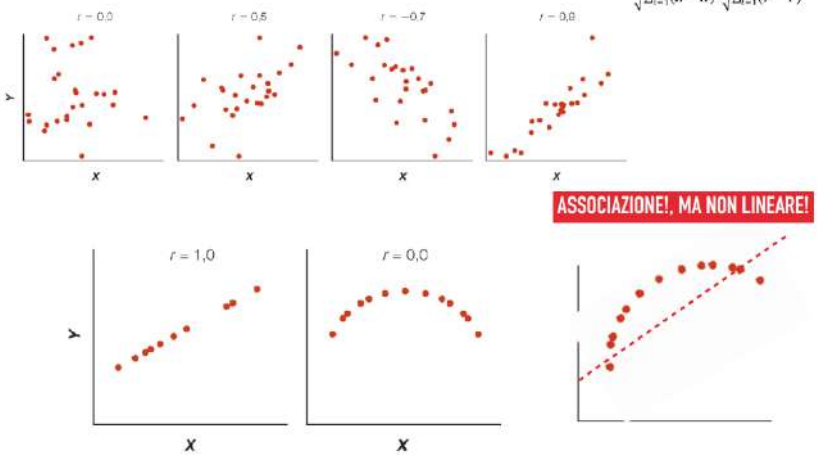
Il coefficiente di correlazione è molto importante e utile in biologia. Questo parametro misura la tendenza di due variabili numeriche (che chiameremo X e Y) a «covariare», cioè, a variare insieme. Il coefficiente r di correlazione viene usato per capire se c'è una regressione lineare tra le variabili prese in considerazione. **Va da -1 a 1;** se è 1, abbiamo una regressione lineare positiva perfetta, se è -1 negativa. se è 0 non c'è regressione lineare.

Il termine al numeratore della formula, detto somma dei prodotti, misura il grado di covariazione (variazione congiunta) degli scarti di X e Y. Lo scarto è la differenza tra un'osservazione e la sua media. Se un'osservazione i si trova al di sopra della media rispetto sia a X sia a Y (quadrante superiore destro della Figura in basso a destra), allora i suoi scarti saranno entrambi positivi e quindi il loro sarà positivo. Se un'osservazione si trova al di sotto della media sia per X sia per Y (quadrante inferiore sinistro), allora entrambi gli scarti saranno negativi e il loro prodotto anche in questo caso sarà positivo. Le osservazioni che si trovano negli altri due quadranti del piano cartesiano hanno uno scarto positivo per una variabile e uno scarto negativo per 1'altra, quindi il loro prodotto degli scarti risulterà negativo. La somma dei prodotti è la somma di tutti i prodotti degli scarti. La somma sarà positiva se la maggior parte delle osservazioni si trova nel quadrante superiore destro e nel quadrante inferiore sinistro del piano. La somma sarà negativa se la maggior parte delle osservazioni si trova nel quadrante superiore sinistro e nel quadrante inferiore destro. Se la dispersione delle osservazioni riempie tutti e quattro i quadranti del piano cartesiano, allora i valori positivi e quelli negativi tenderanno ad annullarsi, fornendo una somma dei prodotti vicina a zero.

Al denominatore nella formula per il calcolo di r compaiono le radici quadrate delle devianze, ossia delle somme dei quadrati per X e per Y.

L'ASSOCIAZIONE TRA VARIABILI NUMERICHE

$$r = \frac{\sum_{i=1}^{n}(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y - \bar{Y})^2}}$$



THM: l'analisi di correlazione funziona solo per associazioni lineari: r=0 non significa non associazione!

COMPRESO TRA -1 E +1. La correlazione massima si ha quando i due punti giacciono sulla retta. **Due variabili possono essere fortemente associate e comunque non avere nessuna correlazione, SE LA RELAZIONE TRA LORO NON è LINEARE**.

INTERVALLO DI CONFIDENZA DEL COEFFICIENTE DI CORRELAZIONE

PERCHE' SERVE CALCOLARE L'INTERVALLO DI CONFIDENZA? **STIMA CAMPIONARIA!!!**

Distribuzione campionaria di *r* non normale

FISHER *trasformazione di r,* $z = 0{,}5 \ln\left(\frac{1+r}{1-r}\right)$; **CAMPIONE**

ζ trasformazione per ρ **POPOLAZIONE**

$$\sigma_z = \sqrt{\frac{1}{n-3}} \qquad z - 1{,}96\sigma_z < \zeta < z + 1{,}96\sigma_z$$

$$r = \frac{e^{2z} - 1}{e^{2z} + 1}$$ **TRASFORMAZIONE INVERSA**

INTERVALLO DI CONFIDENZA $\frac{e^{2zinf} - 1}{e^{2z}inf + 1} < \rho < \frac{e^{2zsup} - 1}{e^{2zsup} + 1}$

THM: IC95% di *r* servono?

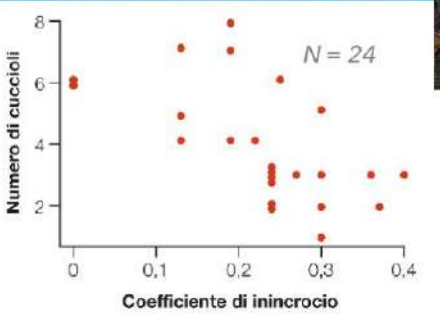
Ovviamente r è una STIMA. Quindi **come tutte le stime ha un INTERVALLO DI CONFIDENZA**.
Una volta trovato l'intervallo di confidenza fatto in modo strano, bisogna riconvertirlo in modo che il PARAMETRO di r, RO, sia davvero quello con le sue giuste unità di misura dentro all'intervallo di confidenza.

VERIFICARE LA SIGNIFICATIVITA' DEL COEFFICIENTE DI CORRELAZIONE

ESERCIZIO

ES: INBREEDING LUPI E SOPRAVVIVENZA INVERNALE DEI CUCCIOLI

inbreeding	pups
0	6
0	6
0.13	7
0.13	5
0.13	4
0.19	8
0.19	7
0.19	4
0.25	6
0.24	3
0.24	3
0.24	3
0.24	3
0.24	2
0.24	2
0.27	3
0.3	5
0.3	3
0.3	2
0.3	1
0.36	3
0.4	3
0.37	2
0.22	4



TEST T STUDENT PER UN CAMPIONE

$$t = \frac{r}{ES_r} \qquad ES_r = \sqrt{\frac{1 - r^2}{n - 2}}$$

$ES_r = 0{,}169$ **df = n -2**

$t = -3{,}60$

$t_{0,05(2),22} = 2{,}075$ H_0: r=0

$t_{0,01(2),22} = 2{,}82$ H_0: r=0

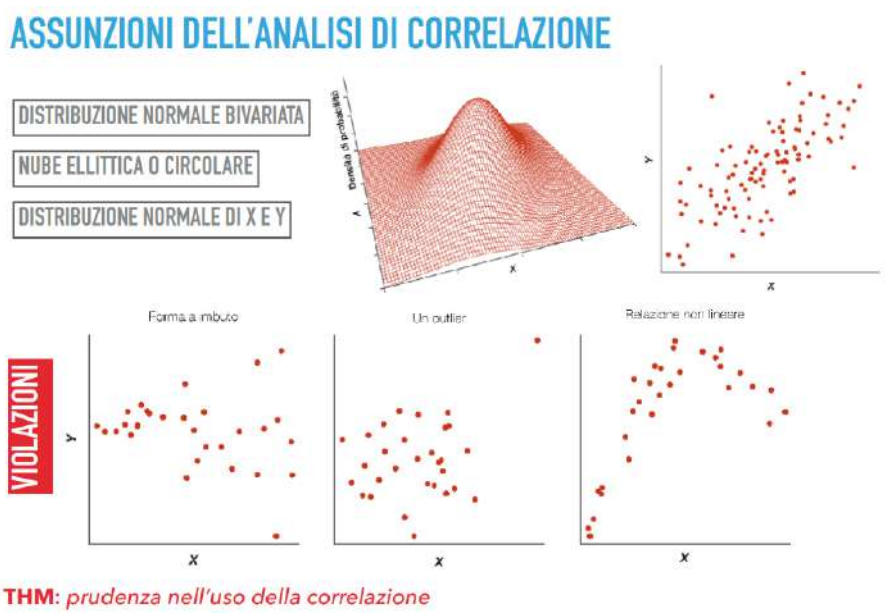
$t_{0,001(2),22} = 3{,}79$ $H_{0;\ 0,001}$: r=0

***r* = ?** **r = -0,608** **p = 0,002**

***p* = ?**

THM: *r* va verificato!!!

IPOTESI NULLA: ASSENZA DI CORRELAZIONE, QUINDI RO=0
Per la verifica delle ipotesi dobbiamo considerare la **STATISTICA t data l'ipotesi nulla**, che si trova con l'errore standard.



distribuzione normale bivariata vuol dire A CAMPANA IN 3D: sia x che y sono normali entrambe.
Per assumere la normale bivariata la cosa migliore da fare è guardare il **DIAGRAMMA DI DISPERSIONE (vedi immagine in basso): se ci sono OUTLIERS, se la relazione tra X e Y non è lineare e se la nube di punti ha una forma a imbuto**.

QUESTO è IL COEFFICIENTE DI PEARSON

ALTERNATIVA: **SPEARMAN**
Caso in cui non serve l'assunzione di normalità. **Si usano i RANGHI**.
Vengono assegnati alle due variabili i ranghi separatamente.
Di solito c'è una formula più facile per calcolare il numeratore perché non ho davvero capito cos'è il rango medio.

THM: il test non parametrico per l'analisi di correlazione lineare si basa sui ranghi

$$\rho_s = 1 - \frac{6\sum_i D_i^2}{N(N^2-1)}$$

dove

$D_i = r_i - s_i$ è la differenza dei ranghi (essendo r_i e s_i rispettivamente il rango della prima variabile e della seconda variabile della i-esima osservazione

N il numero complessivo di osservazioni

È la stessa identica cosa, è solo un pochino più semplice.

ALTERNATIVA ALLA CORRELAZIONE DI PEARSON – SPEARMAN

STATISTICA NON PARAMETRICA **RANGHI!**

$$r_s = \frac{\sum_{i=1}^{n}(R_x - \overline{R_x})(R_y - \overline{R_y})}{\sqrt{\sum_{i=1}^{n}(R_x - \overline{R_x})^2}\sqrt{\sum_{i=1}^{n}(R_y - \overline{R_y})^2}}$$

inbreeding	pups	RANK_inbreed	D_inbreed	Dsq_inbreed	RANK_surv	sqD_surv	D_surv	covar
0	6	1.5	-11.0	121	20.0	56.25	7.5	-82.5
0	6	1.5	-11.0	121	20.0	56.25	7.5	-82.5
0.13	7	4.0	-8.5	72.25	22.5	100	10.0	-85
0.13	5	4.0	-8.5	72.25	17.5	25	5.0	-42.5
0.13	4	4.0	-8.5	72.25	15.0	6.25	2.5	-21.25
0.19	8	7.0	-5.5	30.25	24.0	132.25	11.5	-63.25
0.19	7	7.0	-5.5	30.25	22.5	100	10.0	-55
0.19	4	7.0	-5.5	30.25	15.0	6.25	2.5	-13.75
0.25	6	16.0	3.5	12.25	20.0	56.25	7.5	26.25
0.24	3	12.5	0.0	0	9.5	9	-3.0	0
0.24	3	12.5	0.0	0	9.5	9	-3.0	0
0.24	3	12.5	0.0	0	9.5	9	-3.0	0
0.24	3	12.5	0.0	0	9.5	9	-3.0	0
0.24	2	12.5	0.0	0	3.5	81	-9.0	0
0.24	2	12.5	0.0	0	3.5	81	-9.0	0
0.27	3	17.0	4.5	20.25	9.5	9	-3.0	-13.5
0.3	5	19.5	7.0	49	17.5	25	5.0	35
0.3	3	19.5	7.0	49	9.5	9	-3.0	-21
0.3	2	19.5	7.0	49	3.5	81	-9.0	-63
0.3	1	19.5	7.0	49	1.0	132.25	-11.5	-80.5
0.36	3	22.0	9.5	90.25	9.5	9	-3.0	-28.5
0.4	3	24.0	11.5	132.25	9.5	9	-3.0	-34.5
0.37	2	23.0	10.5	110.25	3.5	81	-9.0	-94.5
0.22	4	9.0	-3.5	12.25	15.0	6.25	2.5	-8.75
		Rango medio	12.5			1098		-728.75
				1123		33.13		1110.43
				33.51				-0.656

ALTERNATIVA ALLA CORRELAZIONE DI PEARSON – SPEARMAN

TEST NON PARAMETRICO **RANGHI!**

$$r_s = \frac{\sum_{i=1}^{n}(R_x - \overline{R_x})(R_y - \overline{R_y})}{\sqrt{\sum_{i=1}^{n}(R_x - \overline{R_x})^2}\sqrt{\sum_{i=1}^{n}(R_y - \overline{R_y})^2}}$$

$r_s = -0{,}656$

THM: non sempre le cose che sembrano difficili lo sono

ALTERNATIVA ALLA CORRELAZIONE DI PEARSON – SPEARMAN

TEST NON PARAMETRICO **RANGHI!**

$$r_s = \frac{\sum_{i=1}^{n}(R_x - \overline{R_x})(R_y - \overline{R_y})}{\sqrt{\sum_{i=1}^{n}(R_x - \overline{R_x})^2}\sqrt{\sum_{i=1}^{n}(R_y - \overline{R_y})^2}}$$

$r_s = -0{,}656$

$$t = \frac{r_s}{ES[r_s]} \qquad ES[r_s] = \sqrt{\frac{1 - r_s^2}{n - 2}}$$

$H_0: \rho_s = 0$?

4,08 0,16

P<0.001 DF=22

$t_{0,05(2),22} = 2{,}075$
$t_{0,01(2),22} = 2{,}82$
$t_{0,001(2),22} = 3{,}79$
$t_{0,0001(2),22} = 4{,}74$

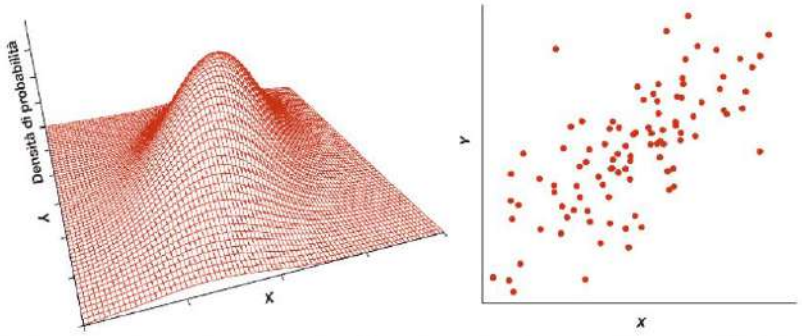
THM: tutte le statistiche devono essere verificate significativamente e spesso con lo stesso test *t* di Student per un campione

Come prima si usa sempre il test t a un campione come riferimento per calcolare l'intervallo di confidenza.

LA REGRESSIONE

X UNA VARIABILE ESPLICATIVA / INDIPENDENTE / PREDITTIVA

Y UNA VARIABILE RISPOSTA / DIPENDENTE / PREDETTA

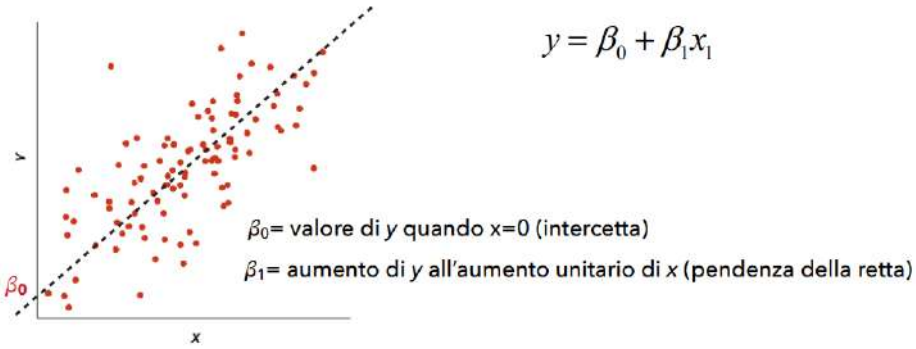


THM: la regressione assume una direzionalita' nella relazione tra due variabili; la correlazione no.

LA REGRESSIONE

X UNA VARIABILE ESPLICATIVA / INDIPENDENTE / PREDITTIVA

Y UNA VARIABILE RISPOSTA / DIPENDENTE / PREDETTA



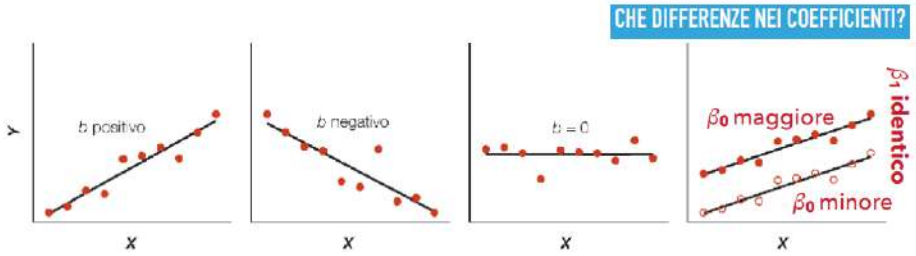
THM: la geometria aiuta

LA REGRESSIONE

X UNA VARIABILE ESPLICATIVA / INDIPENDENTE / PREDITTIVA

Y UNA VARIABILE RISPOSTA / DIPENDENTE / PREDETTA

$y = \beta_0 + \beta_1 x_1$



THM: il coefficiente di regressione unitario fornisce immediatamente importanti informazioni, cosi' come il valore del termine noto.

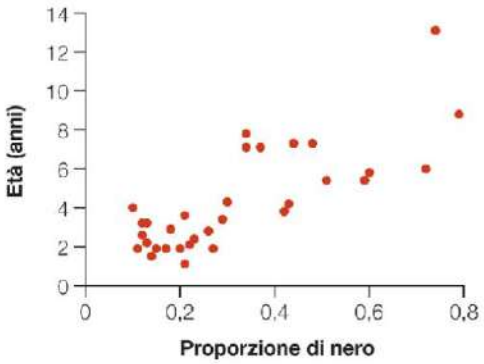
è un metodo per prevedere il valore di una variabile numerica (la variabile di risposta) in base a quello di un'altra variabile numerica (la variabile predittiva).
Il tipo più comune di regressione è la REGRESSIONE LINEARE, in cui si traccia una retta in un diagramma scatter plot per prevedere la variabile di risposta Y a partire da quella esplicativa X. UNA ASSUNZIONE IMPORTANTE è CHE LA RELAZIONE TRA LE DUE VARIABILI SIA DAVVERO LINEARE.

BETA ZERO è l'**INTERCETTA**, che è il VALORE DI Y QUANDO X è ZERO. la sua unità di misura è la stessa di Y.
BETA 1 è invece la **PENDENZA** della retta di regressione, il COEFFICIENTE ANGOLARE: **dice quanto varia Y quando X varia di una unità**.

LA REGRESSIONE

ESERCIZIO ETA' E PIGMENTAZIONE SUL NASO NEI LEONI

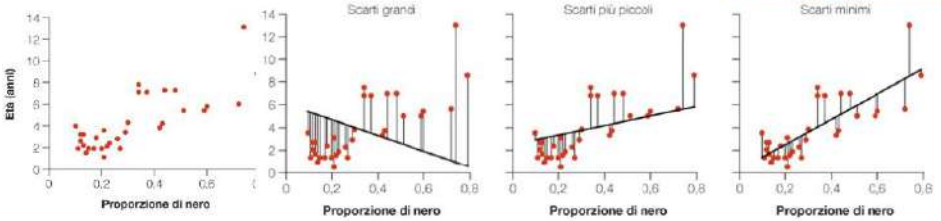
age	prop.black
1.1	0.21
1.5	0.14
1.9	0.11
2.2	0.13
2.6	0.12
3.2	0.13
3.2	0.12
2.9	0.16
2.4	0.23
2.1	0.22
1.9	0.2
1.9	0.17
1.9	0.15
1.9	0.27
2.8	0.26
3.6	0.21
4.3	0.3
3.8	0.42
4.2	0.43
5.4	0.59
5.8	0.6
6	0.72
3.4	0.29
4	0.1
7.3	0.48
7.3	0.44
7.8	0.34
7.1	0.37
7.1	0.34
13.1	0.74
8.8	0.79
5.4	0.51



THM: si parte da una rappresentazione grafica

LA REGRESSIONE – QUALE RETTA?

COME TROVARE I COEFFICIENTI DATA UNA SERIE DI PUNTI?



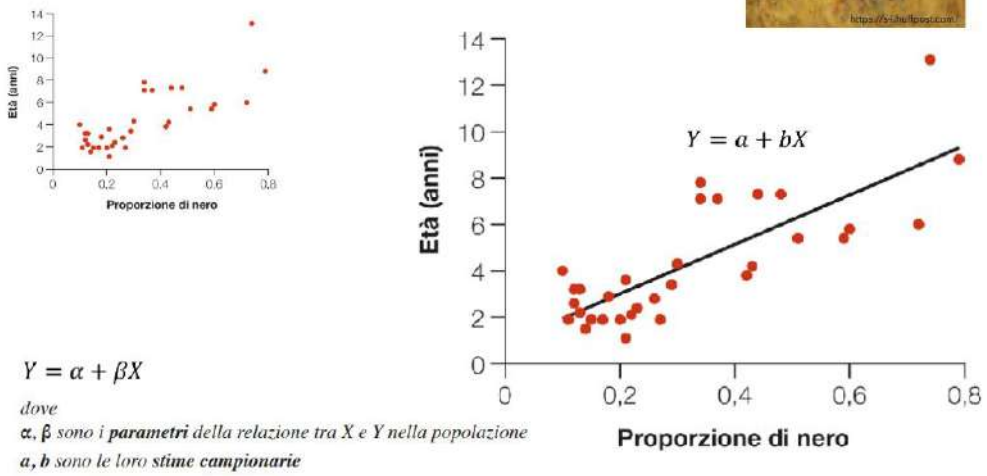
METODO DEI MINIMI QUADRATI

$$Y = a + bX \qquad b = \frac{\sum_1^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_1^n (X_i - \bar{X})^2}$$

la retta passa per il punto $(\bar{X}, \bar{Y})$ $\qquad a = \bar{Y} - b\bar{X} \qquad b = \frac{covarianza(X,Y)}{s_x^2}$

THM: Il metodo dei minimi quadrati e' relativamente semplice

LA REGRESSIONE – QUALE RETTA?



$Y = \alpha + \beta X$

dove

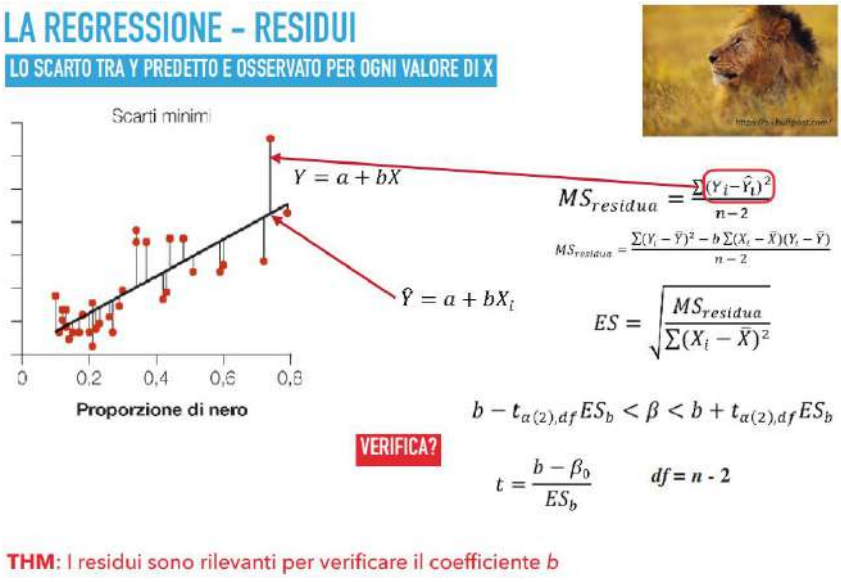
α, β *sono i* ***parametri*** *della relazione tra X e Y nella popolazione*

a, b *sono le loro* ***stime campionarie***

THM: la retta che si ottiene e' CAMPIONARIA

Si usa il computer per calcolare la retta di regressione, comunque si può fare con la calcolatrice con le formule scritte qui: X bar e Y bar sono le medie campionarie delle due variabili X e Y. **Una volta calcolato b, si può trovare l'INTERCETTA a in modo semplice**.
METODO DEI MINIMI QUADRATI.

LA RETTA CHE SI OTTIENE è CAMPIONARIA, CON a E b, NON ALPHA E BETA.



i RESIDUI sono la misura della DISPERSIONE dei punti al di sopra e al di sotto della retta di regressione dei minimi quadrati. Sono molto importanti per valutare l'adattamento della retta di regressione.
PRENDIAMO UNA CERTA X E CALCOLIAMO L'Y PREVISTO CON LA RETTA DI REGRSSIONE CAMPIONARIA CHE ABBIAMO TROVATO. Poi abbiamo anche la **Y del punto vero: possiamo calcolare il MEAN SQUARE DEI RESIDUI**.

Si ha inoltre una STIMA DEL PARAMETRO b, LA PENDENZA. ABBIAMO QUINDI UNA INCERTEZZA ASSOCIATA con il MEAN SQUARE. Possiamo calcolare **l'intervallo di confidenza** in cui si trova il parametro BETA.

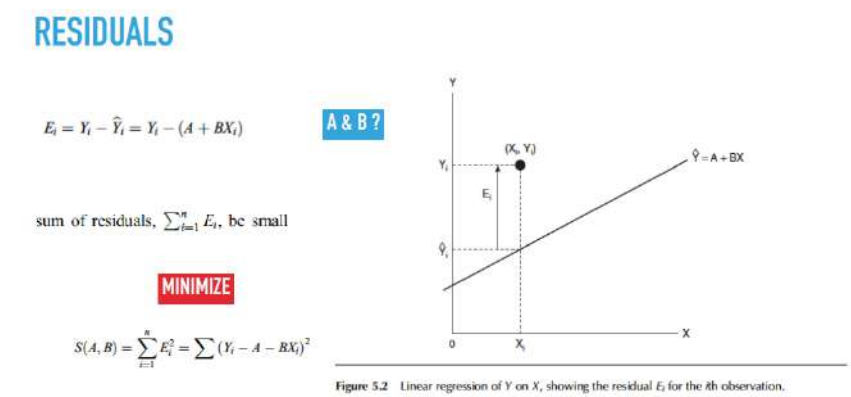
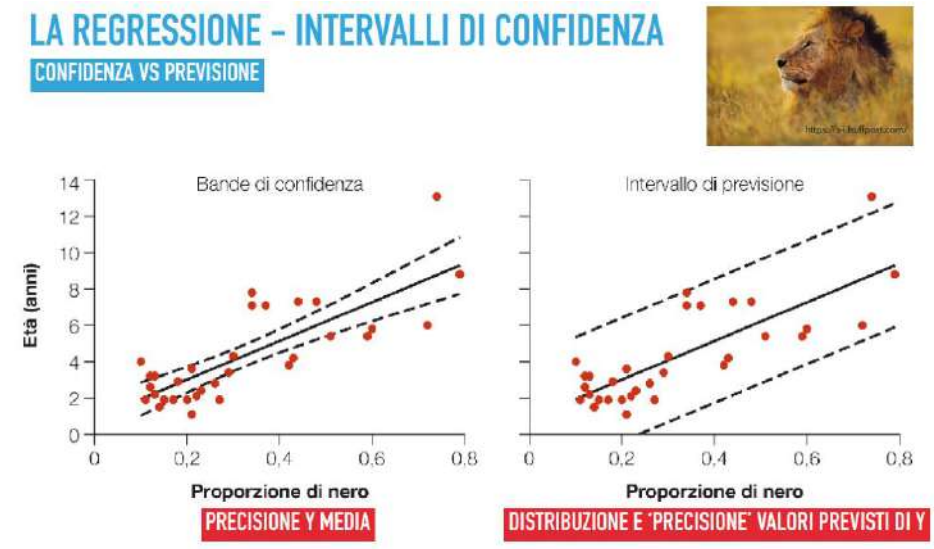


Figure 5.2 Linear regression of Y on X, showing the residual E_i for the ith observation.

In simple linear regression

$$Y_i = A + BX_i + E_i$$

the least-squares coefficients are given by $A = \bar{Y} - B\bar{X}$ and $B = \sum(X_i - \bar{X})(Y_i - \bar{Y}) / \sum(X_i - \bar{X})^2$. The slope coefficient B represents the average change in Y associated with a one-unit increase in X. The intercept A is the fitted value of Y when $X = 0$.



CONFIDENZA O PREVISIONE?
Le bande di CONFIDENZA misurano la precisione di Y media prevista per ogni valore di X. **Queste bande sono quindi PIU' STRETTE IN PRESENZA DI X MEDIO e SI ALLARGANO VERSO L'ESTERNO.**
Gli intervalli di PREVISIONE, invece, misurano la precisione dei singoli valori previsti di Y

per ogni valore di X. Le due linee tratteggiate **sono i limiti superiori e inferiori degli intervalli di previsione per un singolo valore di Y nell'intervallo di valori di X dei dati**.

L'USO DI R^2 PER MISURARE L'ADATTAMENTO DELLA RETTA DI REGRESSIONE AI DATI

È una immediata valutazione del modello, ed è ANALOGO ALL'R^2 DELL'ANALISI DELLA VARIANZA.

LA REGRESSIONE - R^2

$$R^2 = \frac{SS_{regressione}}{SS_{totale}} = \frac{\sum(\hat{Y} - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} \quad mentre\ SS_{residua} = \sum(Y_i - \hat{Y}_i)^2$$

$SS_{totale} = SS_{regressione} + SS_{residua}$

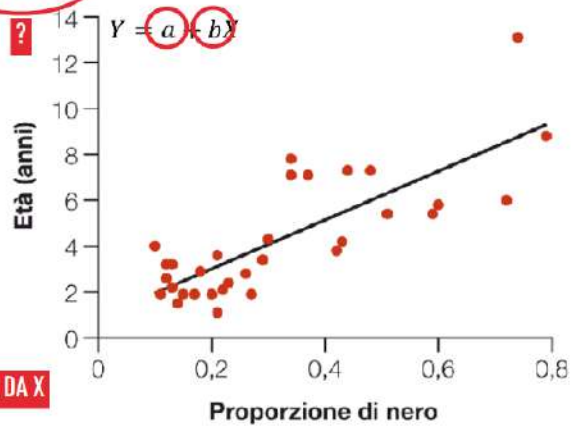
? ?

TEST??? $F = \frac{(SS_{REGRESSIONE}/DF_{MODELLO})}{(SS_{RESIDUA}/DF_{ERRORI})}$

DF??? N-2 (stime parametri: a e b)

$$R^2 = \frac{SS_{regressione}}{SS_{totale}}$$?

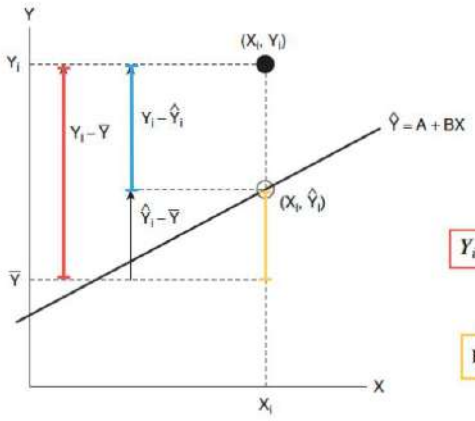
PROPORZIONE DELLA VARIABILITA' DI Y SPIEGATA DA X



THM: R^2 da un'immediata valutazione del modello, che va pero' 'testato'

TOTAL Y VARIANCE DECOMPOSITION - GRAPHICALLY

modo per calcolare graficamente la varianza



$$Y_i - \bar{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})$$

$$RegSS = \sum(\hat{Y}_i - \bar{Y})^2$$

Figure 5.5 Decomposition of the total deviation $Y_i - \bar{Y}$ into components $Y_i - \hat{Y}_i$ and $\hat{Y}_i - \bar{Y}$.

REGRESSION FIT

$$\sum E_i'^2 = \sum (Y_i - \overline{Y})^2$$ *total sum of squares* **TSS** Deviance (variability of y)

-

$$\sum E_i^2 = \sum (Y_i - \hat{Y}_i)^2$$ *residual sum of squares* **RSS**

=

$$\text{RegSS} \equiv \text{TSS} - \text{RSS}$$ **REGSS** REGRESSION SUM OF SQUARES

$$\text{TSS} = \text{RegSS} + \text{RSS}$$

$$r^2 \equiv \frac{\text{RegSS}}{\text{TSS}}$$

$$r = \frac{S_{XY}}{S_X S_Y} = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})^2 \sum (Y_i - \overline{Y})^2}}$$

REGRESSION FIT - CORRELATION

$$r^2 \equiv \frac{\text{RegSS}}{\text{TSS}}$$

$$r = \frac{S_{XY}}{S_X S_Y} = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})^2 \sum (Y_i - \overline{Y})^2}}$$

SYMMETRY

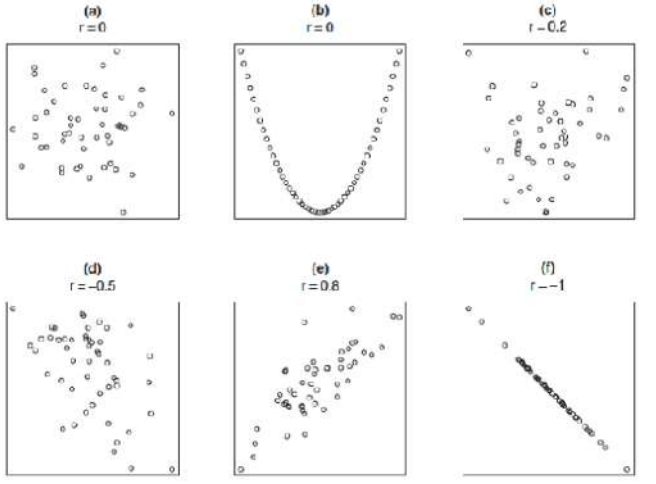


Figure 5.4 Scatterplots illustrating different levels of correlation: $r = 0$ in both (a) and (b), $r = .2$ in (c), $r = -.5$ in (d), $r = .8$ in (e), and $r = -1$ in (f). All the data sets have $n = 50$ observations. Except in panel (b), the data were generated by sampling from bivariate normal distributions.

$$Y_i = A + BX_i + E_i$$
$$= \hat{Y}_i + E_i$$

ASYMMETRY

SYMMETRICAL ONLY IF $R_{(X, Y)} = 1$ OR -1

OBIETTIVI DELLA REGRESSIONE

PREVISIONE
- PREVEDERE Y DATO X
- STIMARE ANDAMENTO

INTERPOLAZIONE O ESTRAPOLAZIONE

SPIEGAZIONE
- COMPRENDERE
- VALUTARE L'EFFETTO DI X SU Y
- FORMULARE/VERIFICARE IPOTESI

$$R^2 = \frac{SS_{regressione}}{SS_{totale}}$$

PROPORZIONE DELLA VARIABILITA' DI Y SPIEGATA DA X

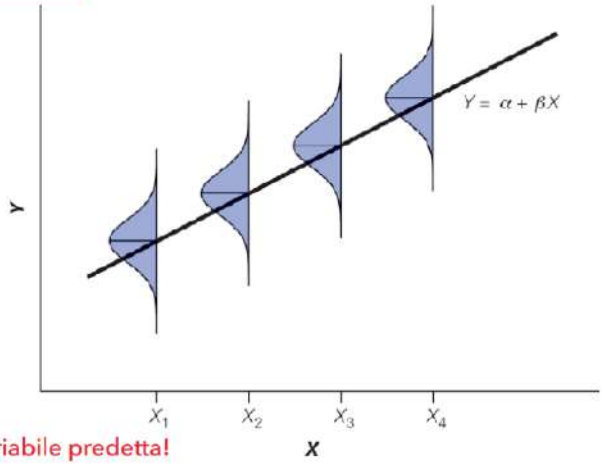
- SE R2 E' ~/=1 → MOLTA VARIABILITA' SPIEGATA
- SE R2 E' << 1 → POCA VARIABILITA' SPIEGATA

THM: un modello si valuta ANCHE sulla base della sua funzione

LA REGRESSIONE - ASSUNZIONI

- PER OGNI X C'E' UNA DISTRIBUZIONE DI Y LA CUI MEDIA GIACE SULLA RETTA (LINEARITA')
- OGNI DISTRIBUZIONE DI Y E' NORMALE E CON VARIANZA SIMILE
- PER OGNI X, Y E' NORMALE E CON VARIANZA SIMILE

NO ASSUNZIONI PER X!!!



THM: le assunzioni sono per la variabile predetta!

ASSUNZIONI DELLA REGRESSIONE: per ogni valore di X esiste una popolazione di possibili valori di la cui media sta sulla retta di regressione (**SI ASSUME CHE LA VERA RELAZIONE SIA LINEARE**).
Per ogni valore di X la distribuzione dei possibili valori di Y è NORMALE.
La varianza dei valori di Y è la stessa per tutti i valori di X.

SIMPLE REGRESSION MODEL – USUAL VIOLATIONS

$$Y_i = \alpha + \beta x_i + \varepsilon_i$$

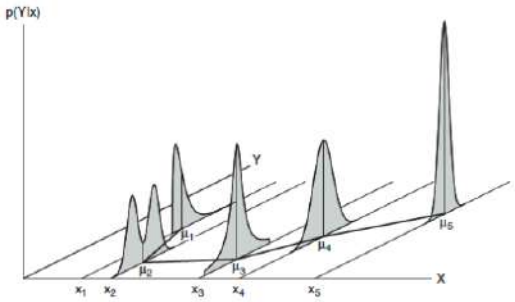


Figure 2.3 Population regression of Y on X. The conditional distribution of Y, $p(Y|x)$, is shown for each of a few values of X. The distribution of Y at $X = x_1$ is positively skewed; at $X = x_2$, it is bimodal; at $X = x_3$, it is heavy tailed; at $X = x_4$, it has greater spread than at $X = x_5$. The conditional means of Y given X, that is, $\mu_1, \ldots, \mu_5$, are not a linear function of X.

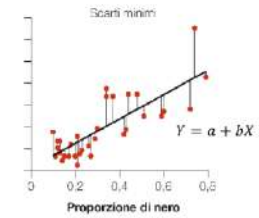
LA REGRESSIONE – RESIDUI X ASSUNZIONI

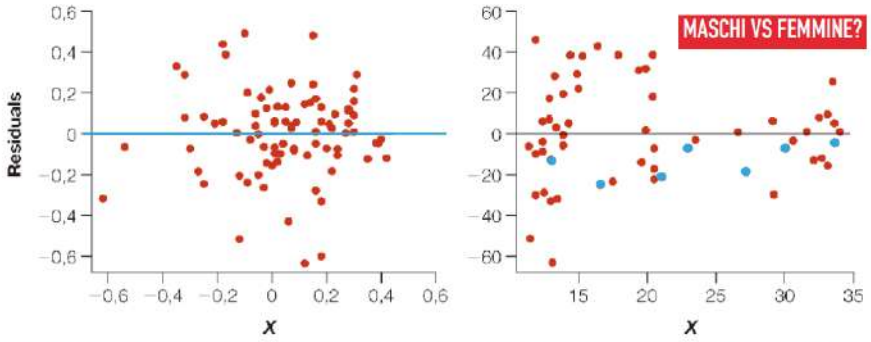
VERIFICARE LE ASSUNZIONI

SIMMETRIA SOPRA-SOTTO

NO CURVATURA

SIMILE DISPERSIONE PER OGNI X

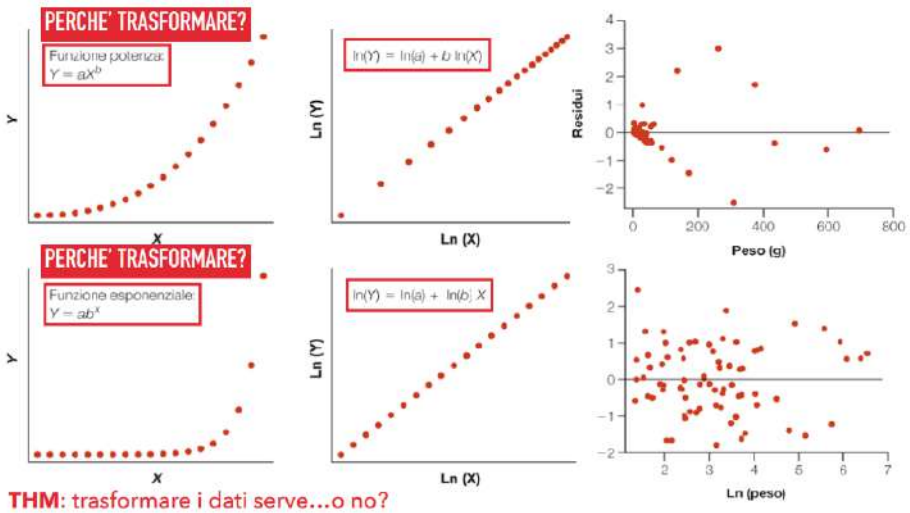




THM: l'analisi dei residui e' rilevante per verificare le assunzioni e identificare anomalie

PER VERIFICARE LE ASSUNZIONI SI FA UNA ANALISI GRAFICA.

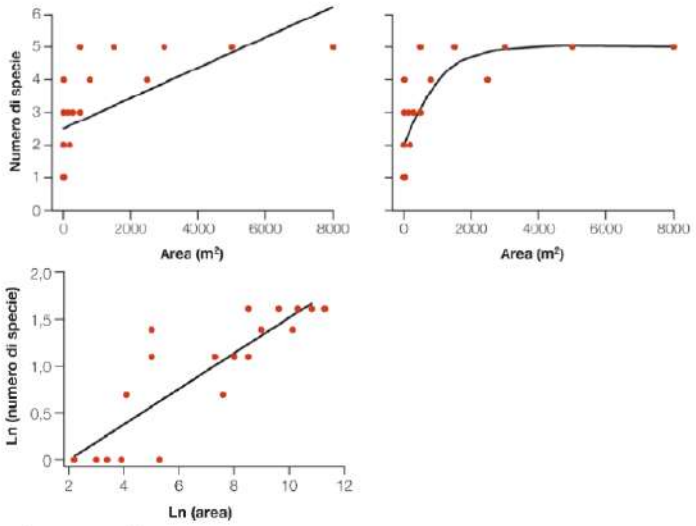
LA REGRESSIONE – TRASFORMAZIONI ?



THM: trasformare i dati serve...o no?

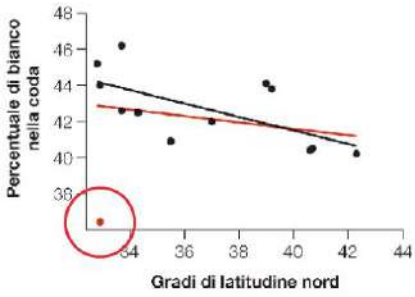
A volte si trasforma anche la variabile per rendere la regressione lineare

LA REGRESSIONE – TRASFORMAZIONI ?



THM: trasformare i dati serve…o no?

LA REGRESSIONE – TRASFORMAZIONI ? OUTLIER

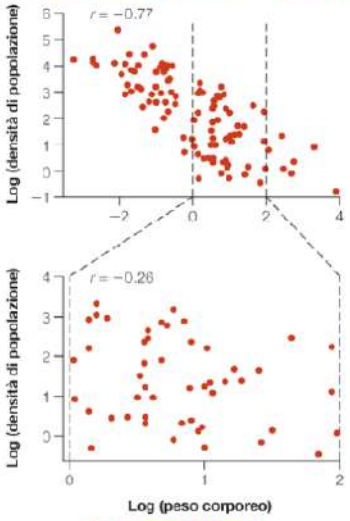


THM: trasformi (se possibile) o cambi approccio

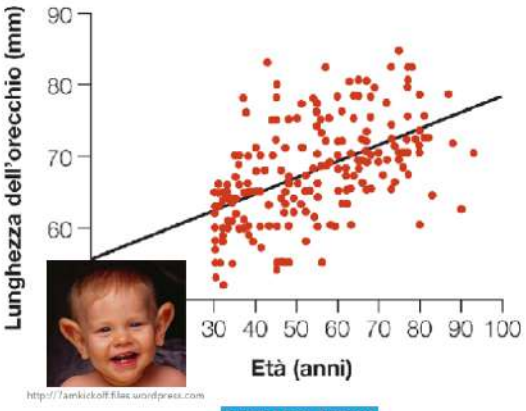
Oltre a creare una distribuzione NON normale di valori di Y per ogni corrispondente valore di X e a violare l'assunto dell'omogeneità delle varianze, GLI OUTLIER INFLUENZANO LE STIME DI PENDENZA E INTERCETTA. **Quindi se è presente un outlier i biologi analizzano QUANTO INFLUENZA I RISULTATI confrontando la retta di regressione con e senza.**

Di solito quelli ai bordi sono molto dannosi. A questo punto si opera con le TRASFORMAZIONI, tipo logaritmica, in modo da avvicinare di più l'outlier, oppure si ricorre ai ranghi.

LA REGRESSIONE - RANGE DI DATI



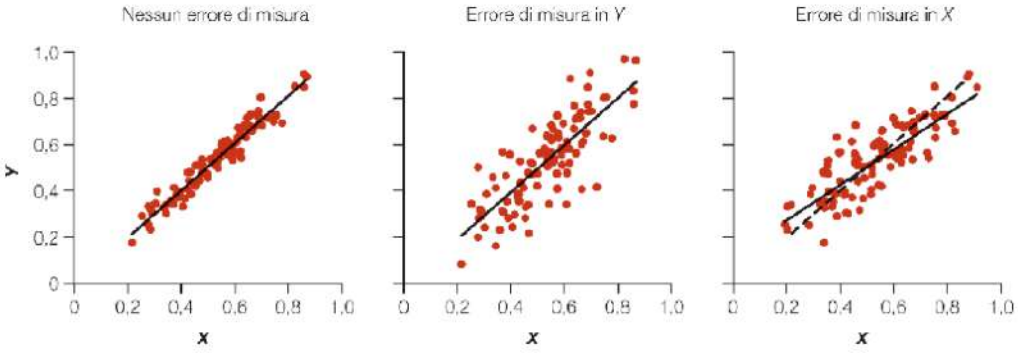
RANGE DI ANALISI



ESTRAPOLAZIONI

THM: attenzione al range di dati considerato

LA REGRESSIONE - ERRORE DI MISURA



WHO IS REGRESSING ON WHOM?

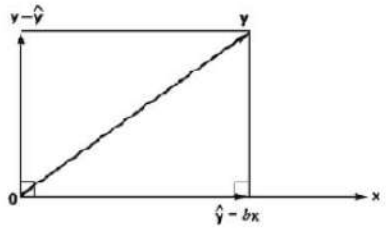


Figure B.11 The orthogonal projection $\hat{\mathbf{y}} = b\mathbf{x}$ of $\mathbf{y}$ onto $\mathbf{x}$.

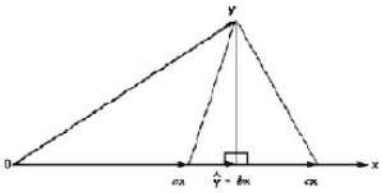
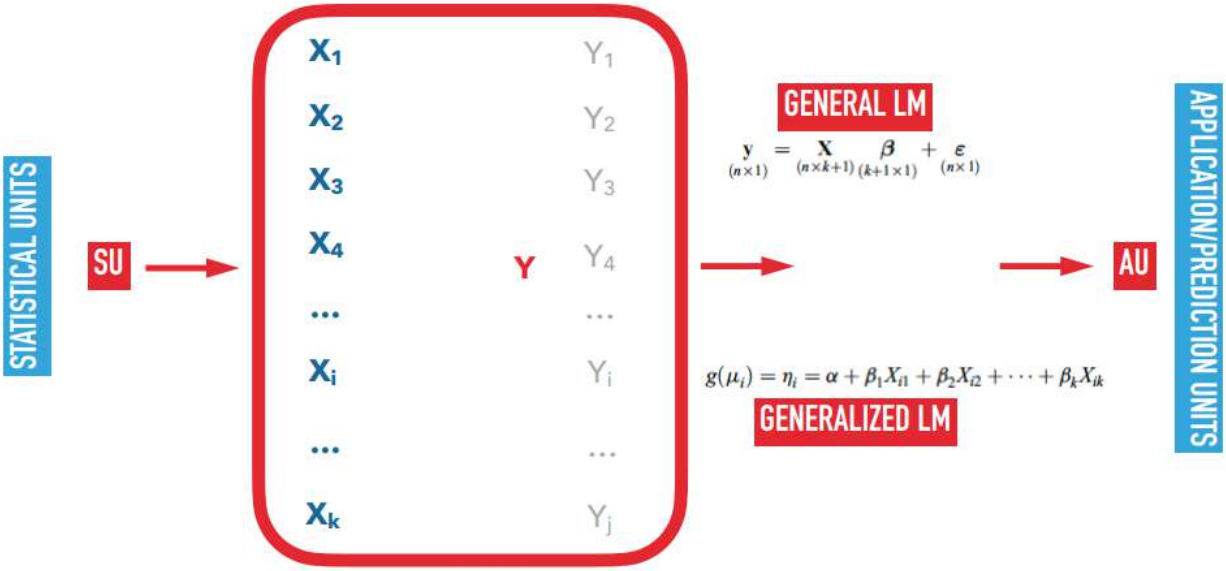


Figure B.12 The orthogonal projection $\hat{\mathbf{y}} = b\mathbf{x}$ is the point along the line spanned by $\mathbf{x}$ that is closest to $\mathbf{y}$

THM: THE REGRESSION IS THE ONE OF Y ON X

MULTIPLE ASSOCIATIONS



MULTIPLE REGRESSION

$$\widehat{Y} = A + B_1X_1 + B_2X_2$$

GEOMETRIC FIGURE?

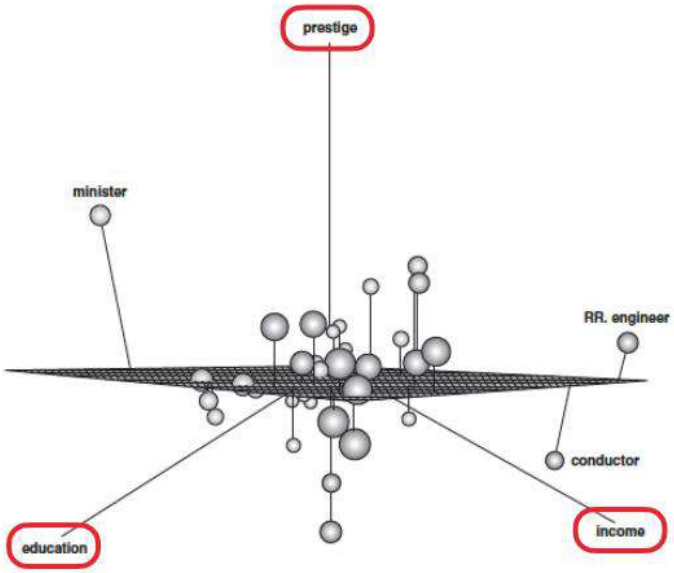


Figure 3.20 Three-dimensional scatterplot for Duncan's occupational prestige data, rotated into an orientation that reveals three unusual observations. From this orientation, the least-squares regression plane, also shown in the plot, is viewed nearly edge on.

MULTIPLE REGRESSION

$$\widehat{Y} = A + B_1X_1 + B_2X_2$$

GEOMETRIC FIGURE?

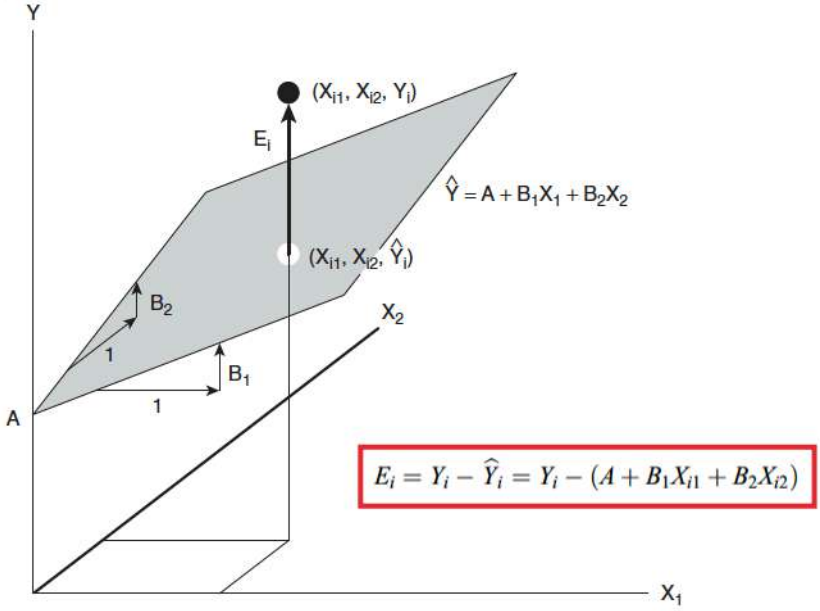
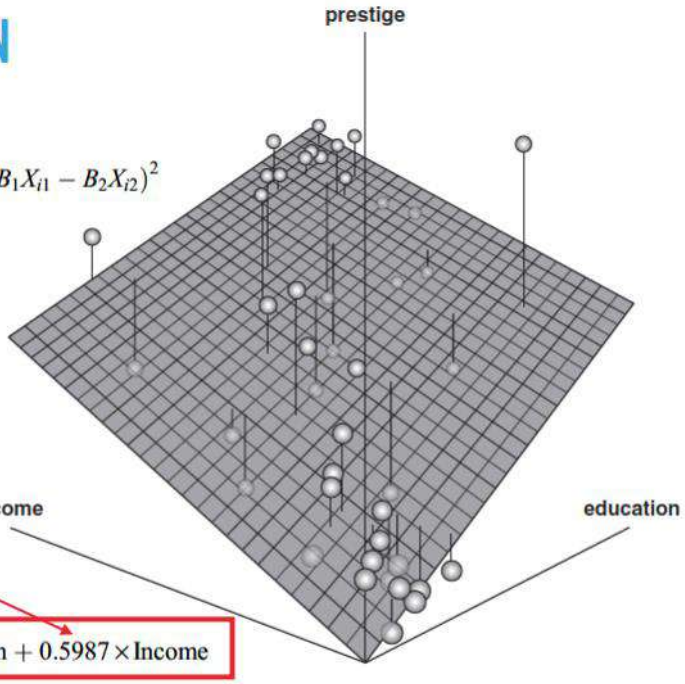


Figure 5.6 The multiple-regression plane, showing the partial slopes B_1 and B_2 and the residual E_i for the ith observation. The white dot in the regression plane represents the fitted value. Compare this graph with Figure 5.2 for simple regression.

MULTIPLE REGRESSION

MINIMISING

$$S(A, B_1, B_2) = \sum E_i^2 = \sum (Y_i - A - B_1 X_{i1} - B_2 X_{i2})^2$$



PARTIAL REGRESSION COEFFICIENTS

$$\widehat{\text{Prestige}} = -6.065 + 0.5458 \times \text{Education} + 0.5987 \times \text{Income}$$

Figure 5.7 The multiple-regression plane in Duncan's regression of prestige on education and income. The two sets of parallel lines on the regression plane represent the partial relationship of prestige to each explanatory variable, holding the other explanatory variable at particular values.

MULTIPLE LINEAR MODEL

$$\begin{aligned} Y_i &= A + B_1 X_{i1} + B_2 X_{i2} + \cdots + B_k X_{ik} + E_i \\ &= \widehat{Y}_i + E_i \end{aligned}$$

K INDEPENDENT VARIABLES

The least-squares coefficients in multiple linear regression

$$Y_i = A + B_1 X_{i1} + B_2 X_{i2} + \cdots + B_k X_{ik} + E_i$$

are found by solving the normal equations for the intercept A and the slope coefficients $B_1, B_2, \ldots, B_k$. The slope coefficient B_1 represents the average change in Y associated with a one-unit increase in X_1 when the other Xs are held constant.

PARTIAL REGRESSION COEFFICIENTS

MULTIPLE LINEAR MODEL

$$Y_i = A + B_1X_{i1} + B_2X_{i2} + \cdots + B_kX_{ik} + E_i$$
$$= \widehat{Y}_i + E_i$$

K INDEPENDENT VARIABLES

The least-squares coefficients in multiple linear regression

$$Y_i = A + B_1X_{i1} + B_2X_{i2} + \cdots + B_kX_{ik} + E_i$$

are found by solving the normal equations for the intercept A and the slope coefficients $B_1, B_2, \ldots, B_k$. The slope coefficient B_1 represents the average change in Y associated with a one-unit increase in X_1 when the other Xs are held constant.

$$S_E = \sqrt{\frac{\sum E_i^2}{n-k-1}}$$

$$R^2 \equiv \frac{\text{RegSS}}{\text{TSS}}$$

The standard error of the regression, $S_E = \sqrt{\sum E_i^2/(n-k-1)}$, gives the "average" size of the regression residuals; the squared multiple correlation, $R^2 = \text{RegSS}/\text{TSS}$, indicates the proportion of the variation in Y that is captured by its linear regression on the Xs.

ADJUSTED SQUARED MULTIPLE CORRELATION COEFFICIENT

$$\widetilde{R}^2 \equiv 1 - \frac{S_E^2}{S_Y^2} = 1 - \frac{\dfrac{\text{RSS}}{n-k-1}}{\dfrac{\text{TSS}}{n-1}}$$

Unless the sample size is very small, however, $\widetilde{R}^2$ will differ little from R^2.

INTERPRETING AND COMPARING MULTIPLE LINEAR MODEL COEFFICIENTS

$$Y_i = A + B_1X_{i1} + B_2X_{i2} + \cdots + B_kX_{ik} + E_i$$

K COEFFICIENTS

The least-squares coefficients in multiple linear regression

$$Y_i = A + B_1X_{i1} + B_2X_{i2} + \cdots + B_kX_{ik} + E_i$$

are found by solving the normal equations for the intercept A and the slope coefficients $B_1, B_2, \ldots, B_k$. The slope coefficient B_1 represents the average change in Y associated with a one-unit increase in X_1 when the other Xs are held constant.

$$\widehat{\text{Prestige}} = -6.794 + 4.187 \times \text{Education} + 0.001314 \times \text{Income} - 0.008905 \times \text{Percent women}$$

Unitary increment in education (in years)
increment in income (1000 $)
Percent of women (in %)

INTERPRETATION?

COMPARING EFFECTS?

DIFFERENT MEASUREMENT UNITS **NOT DIRECTLY COMPARABLE**

e.g. Unitary increment in education (in years)
vs.
increment in income (1000 $)

RESCALING / STANDARDISATION

CORRELAZIONE

IL COEFFICIENTE DI CORRELAZIONE MISURA LA FORZA DELL'ASSOCIAZIONE LINEARE TRA DUE VARIABILI (ASSUNZIONI: normalità bivariata e campionamento rappresentativo)

Cosa vuol dire normale BIVARIATA
Relazione tra X e Y LINEARE
La nube di punti nello scatter plot è circolare o ellittica
Le distribuzioni di frequenza di X e Y sono normali

IL TEST t SERVE PER VERIFICARE L'IPOTESI DI ASSENZA DI CORRELAZIONE LINEARE, e verifica l'ipotesi nulla che il parametro di r sia ZERO
UGUALE PER SPEARMAN MA BASTA CHE CI SIA RELAZIONE LINEARE TRA I RANGHI DI X E Y, non una distribuzione normale bivariata

REGRESSIONE LINEARE

Stima della pendenza della retta (BETA), cioè, trovo b.
ASSUNZIONI: relazione tra X e Y è LINEARE. La distribuzione di Y per ogni valore di X è normale, e la varianza è la stessa per ogni valore di X.
ALTRA ASSUNZIONE è CHE I RESIDUI DEVONO ESSERE SIMMETRICI TRA SOPRA E SOTTO

Perché metto come IPOTESI NULLA che BETA SIA ZERO? Perché voglio capire QUANTO i due fattori che considero sono correlati, e sono TANTO PIU' CORRELATI QUANTO PIU' È PENDENTE LA RETTA, visto che io voglio trovare il principale fattore da cui dipende la mia variabile

R^2: più è alto meglio è: quanto più è alto, quanto più i miei dati sono approssimati dalla retta! Stessa cosa per R^2 del test ANOVA: quanto più è alto, quanto più la variabilità è tra i gruppi e non entro i gruppi.

TEST DI ASSOCIAZIONE: non solo chi quadro e odds ratio, ma anche REGRESSIONE LINEARE E NON LINEARE!!!!